The KK Principle and Rotational Symmetry
(3 June 2020, to appear in *Analytic Philosophy*)

Timothy Williamson

1. *Introduction*

Like other branches of philosophy, epistemology faces a methodological challenge: to make its procedures more robust, so that errors are more easily corrected. The problem is not any general unsoundness in its basic methods, but difficulty in recovering from errors once made.

For example, thought-experimentation is a legitimate and ordinary way of learning what would have been in counterfactual possibilities, as I have argued elsewhere (Williamson 2007). Nevertheless, it is fallible, just as sense-perception, memory, and reasoning are fallible. Although the communal nature of epistemology means that merely idiosyncratic mistakes in designing and conducting a thought-experiment tend to be quickly recognized, that is not the only possible kind of error. A particular thought-experiment may exploit a glitch in human cognitive psychology to which we are all prone. For instance, the set-up may prompt us unconsciously to apply a common heuristic in conditions beyond its range of reliability. How can we guard against the danger of wrongly dismissing a correct theory as refuted because it conflicts with a heuristic in such a case? The answer is not to stop using all cognitive faculties which can lead us astray, for that would leave us with nothing. We need, not ways of never making mistakes, but ways of recovering from our mistakes once made.

We can do better by using more than one method. When diverse methods lead us to the same conclusion, we have a more robust basis for endorsing it. When they lead to conflicting conclusions, we have a warning sign that something has gone wrong. Each method acts as a potential corrective to the others. For the method of thought-experimentation, an appropriate corrective is the method of formal model-building—and *vice versa* (Williamson 2017). Thus formal models provide independent confirmation for the morals of Gettier thought-experiments (Williamson 2013).

The issue of robustness also arises *within* the model-building approach. Some features of a model may result from arbitrary or at least unforced choices; a better model may avoid them. We can have more confidence in features which are somehow *persistent*: more specifically, features which provably follow from appropriate general constraints on models. This paper exemplifies that approach, applying it to counter-models to the controversial ‘KK’ principle that whenever you know *p*, you also know that you know *p*. It is to be hoped that a similar approach can also be applied to other questions in epistemology.

2. *The KK principle and model-building*

The counter-models to KK are inspired by cases of the following sort. Notoriously, perceptual indiscriminability is non-transitive. Imagine a long line of people, arranged in height from the tallest to the shortest, standing some distance away. They form a sorites-like series. For all you can see of any two successive members, they are exactly the same height; thus they are visually indiscriminable in height for you (in those circumstances). Yet the first and the last members are *not* visually indiscriminable in height for you: you can easily see that the former is much taller than the latter. Thus, on pain of contradiction, for some three people A, B, and C in the series, A is visually indiscriminable in height from B, and B is visually indiscriminable in height from C, but A is visually discriminable in height from C.

On the intended interpretation of standard possible world models of epistemic logic, a world *x* is accessible from a world *w* just in case, for all one knows in *w*, one is in *x*—in other words, whatever one knows in *w* is true in *x*. Thus accessibility is understood as a kind of indiscriminability. Since knowledge is factive, whatever one knows in *w* is true in *w*, so accessibility is reflexive: no world can be discriminated from itself. One counts as knowing a proposition in *w* just in case the proposition is true in every world accessible from *w*. As is well known, the transitivity of accessibility corresponds to the validity of the KK principle.[1] An obvious thought is therefore to use instances of the non-transitivity of perceptual indiscriminability to construct natural, somewhat realistic models of epistemic logic which falsify the KK principle, because their accessibility relation is non-transitive. The task is not trivial, since non-transitive indiscriminability between objects of perception has to be leveraged into non-transitive indiscriminability between worlds, without sacrificing too much naturalness or realism. Nevertheless, it can be done (Williamson 1992, 2000).

One bonus of working with indiscriminability is that the counterexamples also work against a more defensible weakening of KK which merely requires that whenever you know *p*, you are at least *in a position to* know that you know *p*. For indiscrimin*ability* is not just incompatible with knowing the relevant difference in the relevant way; it is also incompatible with being in a position to know the difference in that way.

Of course, standard epistemic logic is widely regarded as quite unrealistic. Specifically, its models validate an unrestricted principle of multi-premise closure for knowledge: whatever truths you know, you also know whatever truths follow from them (in the epistemic logic)—with no qualifying condition to the effect that you have carried out the deduction, or anything like that. Such logical omniscience seems way beyond the computational powers of even the cleverest mortals. One response is to try to gloss ‘know’ in some way that automatically validates unrestricted multi-premise closure, though that risks changing the meaning of ‘KK’, and leaving the original principle intact. Another response is to isolate just the required bits of the model and present them as an informal description of an ordinary scenario, without commitment to an unrestricted closure principle. A third attitude is just to admit that the model characterizes a drastically idealized agent, but to argue that if even such an idealized agent is susceptible to failures of KK, we can hardly hope to do better. Satisfying KK is not a booby prize for those who are bad at logic.

Recently, there has been some pushback in defence of KK, with its epicentre at MIT (Greco 2014, Stalnaker 2015, Das and Salow 2018, Goodman and Salow 2018, Dorst 2019, but also McHugh 2010). Notably, much of this movement has developed within an epistemological approach not unsympathetic in spirit to that of the original critique of KK: a broadly externalist, reliabilist conception of knowledge, a willingness to treat knowledge on its own terms rather than reduce it to some sort of belief with privileges, an openness to applying the techniques of formal epistemology.

A generic feature of the methodology of formal model-building is that when one argues for the possibility of a phenomenon (such as KK failure) on the grounds that simple models of some type of situation predict it, someone can counter by building more complex models of such situations which do not predict the phenomenon. That has in effect happened for KK failure. Of course, if the extra complications look gerrymandered just to avoid the prediction, the counter is open to the objection that it is *ad hoc*, and not scientifically respectable: that has to be judged case by case. In any case, it would be nice to *generalize* the original argument for the possibility of the phenomenon, by showing the prediction to be robust because it follows from quite general features of the model, which should be preserved even as further dimensions of complexity are added: it is no mere artefact of the extreme simplicity of the original model. The next three sections do that for one type of counter-model to KK. The final section uses that result to reflect on recent defences of KK.

3. *Unmarked clocks*

Clock models were originally proposed to argue for a phenomenon much more drastic than KK failure: the possibility of knowing *p* while it is almost certain on one's own present evidence (or knowledge) that one does not know *p* (Williamson 2011, 2014). The failure of KK is in effect a corollary of that phenomenon. However, working with a probability distribution requires specifying much more structure over the model, in a way which in this case is hard to generalize in an adequately motivated way to the desired extent, especially for models with infinitely many worlds. We therefore aim at a more modest target, simplifying matters by considering the original model just as a proposed counter-model to KK.

Imagine looking from some distance at an unmarked circular clock-face with just one rotating hand. That is your only source of knowledge of the hand's position (for example, 3 o'clock). Your question is where the hand is pointing (what time it is pointing at), not whether the clock is accurate. We treat the non-epistemic possibilities as simply positions, or points on a circle. By looking, you gain *some* knowledge of the hand's position—you can rule out some of those possibilities—but you do not gain *full* knowledge—you cannot rule out all of them except the actual one.

The original model was discrete, with finitely many equally spaced positions for the hand to be at, in order to avoid tractable but messy complications with probability distributions over an infinite probability space. For brevity, we simply write 'positions' for the positions available for the hand to be at. Since we have put probability aside here, infinite models raise no special problem. We can have a continuous model, with positions on a segment of the circle ordered like the real numbers in a bounded interval, which is geometrically more natural. If we want to denote each position with a finite string of symbols from a finite alphabet, we can make the set Θ of positions countable, with positions on a segment ordered like the rational numbers in a bounded interval. The choice between integer-valued, real-valued, and rational-valued positions makes little difference to the argument below (though it would need adjustment if we allowed infinitesimal distances). Irrespective of that choice, we describe Θ as a circle.

What matters is that the positions are too close together for the observer to see exactly which one the hand is at, and that they are evenly spaced, so that Θ has rotational and reflective symmetry, like a circle. Thus for any positions θ_1 and θ_2, some structure-preserving mappings of Θ onto itself map θ_1 to θ_2. The observer's powers of visual discrimination are assumed to show the same symmetries. They are just as good around θ_1 as they are around θ_2, and just as good clockwise as anti-clockwise. Whether the observer can discriminate θ_1 from θ_2 depends only on the angle they subtend at the centre of the circle. In this simple case, the indiscriminability relation between positions constitutes the accessibility relation between worlds, for purposes of epistemic logic.

Strictly speaking, in the terminology of epistemic and modal logic, this set-up is a *frame* rather than a *model*, because it specifies a set of worlds and an accessibility relation over that set, but no specific propositions to interpret the atomic formulas of a formal language for epistemic logic. No such formal language was needed for the purposes at hand. To avoid confusion, we respect the distinction between frames and models in what follows. A result in section 4 is best articulated with reference to a formal language.

The original frame is very simple. The worlds are simply the positions themselves. The worlds accessible from a given world, in effect the positions indiscriminable from the given position, form an interval of constant length centred on the given position. For ease of visualization, we may assume that the interval occupies a comparatively small proportion of the circle. Start from a position θ_1; let θ_2 be a position within the interval of indiscriminability around θ_1 but near its edge going clockwise; let θ_3 be a position within the interval of indiscriminability around θ_2 but near its edge going clockwise. Then θ_3 is not within the interval of indiscriminability around θ_1. In other words, θ_2 is accessible from θ_1, and θ_3 is accessible from θ_2, but θ_3 is not accessible from θ_1. Thus accessibility is non-transitive, so the KK principle fails in the frame. To be more specific, let X be the proposition true at just the worlds accessible from θ_1; thus X is the strongest truth one knows at θ_1—at θ_1, one knows X, and X entails everything else one knows. Then as an observer in θ_1, one knows X but does not know that one knows X.

Although the original frame is very natural, it involves some drastic simplifications, on top of those already built into the standard semantic framework for epistemic logic. By identifying worlds with positions, it restricts the observer's ignorance of epistemic matters to ignorance induced by ignorance of non-epistemic matters. In the frame, no two epistemic possibilities alike in the position of the hand differ in what the observer knows. For example, the observer in effect knows the exact length of the interval of indiscriminability, since it is constant from one world to another. By contrast, in real life, observers are typically ignorant of some such structural features of their own cognitive powers.

Another special feature of the original frame is that its accessibility relation is symmetric: whenever θ_1 is indiscriminable from θ_2, θ_2 is indiscriminable from θ_1. Yet non-symmetric accessibility relations are needed in epistemology, for example, to model a sceptical scenario without letting scepticism infect the corresponding non-sceptical scenario. In the bad case, for all one knows one is in the good case, but in the good case one knows things incompatible with being in the bad case.

Thus the worry arises that introducing extra structure into the frame to make it more realistic in such respects may somehow disrupt the counterexamples to KK. To counter that worry, we need to generalize. We do so by considering all frames which generalize simple clock frames in the following way.

We still have a set of worlds W, but we no longer equate it with the circle Θ. However, we can map W onto Θ: for each world *w*, [*w*] is the position of the hand in *w*. As usual, propositions are equated with subsets of W. For any position θ, P_θ is the proposition that the hand is at θ; in other words, $P_\theta = \{w \in W: [w] = \theta\}$. The frame encodes knowledge in the usual way, with a dedicated accessibility relation R. For any proposition X, KX is the proposition that (as the observer) one knows X; as usual, KX is true if and only if X is true in all accessible worlds, so $KX = \{w \in W: \forall x(\text{if } Rwx, x \in X)\}$. But R can no longer be defined simply by angular distance, since that distance may not determine all the differences between worlds. When $[w] = [x]$ but $w \neq x$, whether *x* is accessible from *w* may depend on epistemic differences between *w* and *x*, or even on non-epistemic differences, for we allow a frame to have additional structure beyond W and R, although W and R are our ultimate concern. We do *not* require R to be symmetric.

The main structural constraint on the frames is *rotational symmetry*: we require them to look structurally the same when ‘rotated’ through any angle. The frame itself gives no structurally privilege to one position over any other. Since the frames are not ordinary geometrical objects, we must be more precise about just what is required. We postulate that for each rotation *r* of the circle Θ, there is a corresponding *automorphism r** of the frame, a structure-preserving bijection of W onto itself, which can be pictured as rotating the frame in a manner analogous to that in which *r* rotates the circle. Crudely, we can think of $r^*(w)$ as a world just like *w* but for the rotation of the circle through *r*. We can state the main effect of r^* in terms of *r* thus, for all rotations *r*:

(1a) $[r^*(w)] = r([w])$

In other words, the hand’s position in the result of applying the automorphism to a world is the result of applying the corresponding rotation to the hand’s position in the original world.

For present purposes, rotating clockwise through 90° counts as the same rotation as rotating anti-clockwise through 270°, rotating through 360° counts as the same rotation as rotating through 0°, and so on, because they have the same effect. Rotations are *functions*, not temporal processes.

Rotations of the circle induce rotational automorphisms of the frame <W, R> in a natural way, so that the rotational automorphisms inherit much—though not all—of the structure of the rotations themselves (section 4 will explain in detail one general way for this to happen). In particular, this concerns their algebraic structure. The rotations of the circle form a *group* in the mathematical sense, under the operation of functional composition. There is an identity rotation $\underline{1}$ which leaves everything where it was: $\underline{1}(\theta) = \theta$ for every position θ. Every rotation r has a two-sided inverse rotation r^{-1} which cancels out r: $r^{-1}(r(\theta)) = r(r^{-1}(\theta)) = \theta$. Rotations can be composed: applying first r_1 and then r_2 is equivalent to applying the rotation r_2r_1: $r_2r_1(\theta) = r_2(r_1(\theta))$. Thus $\underline{1}r = r\underline{1} = r$ and $rr^{-1} = r^{-1}r = \underline{1}$. The group of rotations of the circle induces a corresponding group of automorphisms of the frame. We call such automorphisms r^* of the frame *rotational automorphisms*. The frame may have non-rotational automorphisms too, but we are less interested in them. The automorphisms of any structure automatically form a group, under composition (which is always associative in the mathematical sense), and in this case the rotational automorphisms form a subgroup of that group. Its group operations are induced by those of the group of rotations. For any rotations r_1 and r_2, the rotational automorphism $(r_2r_1)^*$ is just the composition of r_1^* and r_2^*: $(r_2r_1)^*(w) = r_2^*(r_1^*(w))$. Consequently, the rotational automorphism $\underline{1}^*$ is the identity automorphism, for $\underline{1}^*r^* = (\underline{1}r)^* = r^*$ for any rotation r. Similarly, r^{-1*} is the two-sided inverse of r^*, for $r^{-1*}r^* = (r^{-1}r)^* = \underline{1}^* = (rr^{-1})^* = r^*r^{-1*}$.

Since the rotational automorphisms form a subgroup, we can define an equivalence relation E over W by setting Ewx just in case for some rotation r, $r^*(w) = x$. E is reflexive because the identity is a rotational automorphism; it is symmetric because rotational automorphisms are closed under inverses; it is transitive because they are closed under composition. Thus E partitions W into mutually exclusive, jointly exhaustive subsets, called *orbits*. Studying epistemic phenomena on a fixed orbit will turn out fruitful.

A distinctive feature of the circle is that for any positions η and θ, exactly one rotation r is such that $r(\eta) = \theta$. The epistemic frame need not inherit that feature. Typically, some worlds differ from each other structurally with respect to R, so that *no* automorphism of the frame maps one to the other. Then W divides into more than one orbit. However, the frame does inherit the feature of *uniqueness* from the circle: *at most* one rotational automorphism maps a world w to a world x. For if $r_1^*(w) = r_2^*(w)$ then $[r_1^*(w)] = [r_2^*(w)]$, so by (1a) $r_1([w]) = r_2([w])$, so $r_1 = r_2$ by uniqueness for the circle. As a corollary, the * function is one-one: whenever $r_1^* = r_2^*$, $r_1 = r_2$, so distinct rotations never induce the same rotational automorphism. This also means that for any worlds w and x in the same orbit, there is a unique rotation r of the circle such that $r^*(w) = x$. We can therefore use the size of the angle (in degrees) subtended by positions θ and $r(\theta)$ at the centre of the circle (which is independent of θ) as a natural numerical measure of the 'distance' $|w, x|$ between the worlds w and x. Of course, this measure is defined only for worlds in the same orbit.

These ideas enable us to state a natural monotonicity condition on epistemic accessibility, for all worlds w, x, and y:

(1b) If w, x, and y are in the same orbit, and $|w, x| \geq |w, y|$, then Rwx only if Rwy

For when the worlds differ only by rotation, and y is at least as close to w as x is, then x is indiscriminable from w only if y is indiscriminable from w. In that way, discriminability of worlds from a given world increases with distance from that world.

A third constraint is this. For any rotation r, since r^* is an automorphism, it preserves the structure of the frame, and since R is a component of that frame, r^* preserves R, in the sense that for all worlds w and x:

(1c) $Rr^*(w)r^*(x)$ if and only if Rwx

In other words, the accessibility relation is invariant under rotational automorphisms: they preserve the epistemic structure of the frame. This is just part of what is meant by saying that r^* is an automorphism of the frame.

If the frame has more structure on W, beyond R, we should add analogues of (1c) for the additional properties, relations, or functions. Again, these invariance principles only mean that the frame does not introduce arbitrary deviations from the rotational symmetry of the underlying set-up.

Crucially, the postulated symmetries are of the frame as a whole, *not* within each world in the frame. The most obvious reason for this is that in each world w the hand is at a specific position $[w]$, which breaks the *intra*-world symmetry between $[w]$ and any other position $r([w])$. Rotational symmetry is *inter*-world symmetry: in another world $r^*(w)$ the hand is at $r([w])$, as in (1a).

More subtly, a psychological bias may break the intra-world symmetry: for example, its effect may be that one is much better at discriminating around 6 o'clock and 12 o'clock than around 3 o'clock and 9 o'clock. That is quite consistent with inter-world rotational symmetry at the level of the frame, which merely implies that the frame contains some *other* world in which one has a psychological bias whose effect is that one is correspondingly better at discriminating around 3 o'clock and 9 o'clock than around 6 o'clock and 12 o'clock.

We do *not* assume that $r^*(w)$ is causally the closest world to w in which the position of the hand is $r([w])$. In the case just considered, the psychological bias may be quite deeply rooted, and hard to 'rotate'. Thus if w is the world in which the hand is at 6 o'clock and one's bias favours 6 o'clock and 12 o'clock, causally the closest world in which the hand is at 9 o'clock may be one in which one's bias still favours 6 o'clock and 12 o'clock, not one in which it has been rotated to favour 3 o'clock and 9 o'clock, whereas the world $r^*(w)$ is of the latter kind.

Something is wrong with any view of knowledge which cannot handle such naturally symmetric frames.

The remaining two constraints are stated with respect to a fixed world z, which we informally envisage as a typical instance of the more 'normal' or 'realistic' worlds in the frame, while allowing that the frame may also contain less realistic, more abnormal worlds, perhaps as merely epistemic possibilities, or something even more distant than that. Thus the constraints on z are *not* constraints on all worlds in the frame.

The fourth constraint just says that the observer in z has non-trivial knowledge about the position of the hand. Recall that the negation of a proposition X in the frame is its complement in W, W−X; thus $K(W-P_\theta)$ is the proposition that the observer knows that the hand is *not* at the position θ:

(1d) For some $\theta \in \Theta$, $z \in K(W-P_\theta)$

In particular, since K is factive, $z \in W{-}P_\theta$, so $[z] \neq \theta$. By (1d), in z one can rule out *some* candidate positions for the hand; looking at the clock was not a complete waste of time. The natural alternative to (1d) would be some form of scepticism about the external world.

The final constraint means that in z the observer's powers of discrimination are not perfect over its orbit; some rotation other than the identity $\underline{1}$ is too small to detect:

(1e) For some rotation $s \neq \underline{1}$, $Rzs^*(z)$

In other words, as the observer in z, one cannot discriminate the world one is in from the result of slightly rotating it. By condition (1c), $s^*(z)$ is a world with the same epistemic structure as z: it differs from z only in ways consequent on a slight difference in the position of the hand. This is consistent with there being other worlds in the frame where the hand has the same position as in $s^*(z)$ but the structure of the observer's knowledge differs markedly from that in z: the observer in z may be able to rule out being in one of the latter worlds, but cannot rule out being in $s^*(z)$. Notably, (1e) allows the observer's powers of discrimination to be greater in some worlds than in others, but then the former cannot be rotated to the latter. For reasons already explained, $s^*(z)$ may not be causally the closest world to z in which the position of the hand is $s([z])$; the latter world may be one in which some psychological bias in z has not been rotated by the analogue of s. Nevertheless, epistemically, for a small enough rotation s, the observer in z cannot discriminate their world from one just like it but correspondingly slightly rotated by s^*. The natural alternative to (1e) would be an implausible kind of omniscience about the relevant aspect of the world, for if (1e) fails, the observer in z—as well as being logically omniscient—can in effect discriminate their position in the world from an arbitrarily close and structurally exactly similar position, differing only in ways correlated with an arbitrarily slight difference in the real and apparent positions of the hand. Neither normal vision nor normal introspection has so perfect a power of discrimination.

Given (1a)-(1e), we can prove that R is non-transitive. For let z, θ, and s verify (1d)-(1e). Then for some small enough rotation r and some natural number $n \geq 1$, $\theta = r^n([z])$ (where r^n is n iterations of r) and $|z, s^*(z)| \geq |z, r^*(z)|$ (we assume the Archimedean property of the standard real numbers: the argument would not work in its present form if $|z, s^*(z)|$ were infinitesimal). Then $Rzr^*(z)$ by (1b). So, for each $m < n$, by (1c) $Rr^{*m}(z)r^{*m}(r(z))$, i.e. $Rr^{*m}(z)r^{*(m+1)}(z))$. Suppose that R is transitive. Then $Rr^{*0}(z)r^{*n}(z)$, i.e. $Rzr^{*n}(z)$. Thus, by (1d) and the definition of K, $r^{*n}(z) \in W{-}P_\theta$. Hence, by definition of P_θ, $[r^{*n}(z)] \neq \theta$. But $[r^{*n}(z)] = r^n([z])$ by n applications of (1a), and r was chosen such that $\theta = r^n([z])$, so $[r^{*n}(z)] = \theta$. This is a contradiction. Thus R is non-transitive. QED.

By similar reasoning, without assuming that R is transitive, we can show that $n \geq 1$ and $K^n(W{-}P_\theta)$ is false at z (where K^n is n iterations of K). But $K(W{-}P_\theta)$ is true at z by (1d). Thus, for some $m \geq 1$, $K^m(W{-}P_\theta)$ is true at z while $K^{m+1}(W{-}P_\theta)$ is false at z. In other words, $K(K^{m-1}(W{-}P_\theta))$ is true at z while $KK(K^{m-1}{-}P_\theta))$ is false at z, so in the frame an instance of the KK principle fails at the realistic world z itself, not just at some 'far-out' world.

The result shows that arguing against KK from the original frame did not crucially depend on its very simple structure. A version of the argument goes through for a wide range of much richer and more realistic frames. What it takes is just rotational symmetry combined with the idea that, in the envisaged circumstances, perception provides some knowledge but not complete knowledge of the hand's position (on the current orbit). Although rotational symmetry is of course a simplification, rejecting it looks like an *ad hoc* move of the most scientifically unproductive kind.

4. *Combining unmarked clocks with other frames*

To see how much the constraints (1a)-(1e) allow, we can construct a wide range of epistemic frames satisfying them all. Indeed, this section establishes a more precise result: if a formula in a standard language of single-agent epistemic logic is invalid on some reflexive frame, it is invalid on some reflexive frame satisfying (1a)-(1e). In that sense, the constraints are compatible with a wide range of basic epistemic phenomena.

We start by explaining the syntactic and semantic terminology. The standard language of single-agent epistemic logic is just the result of adding a monadic sentential operator *K* (for what the agent knows) to a standard language of non-modal propositional logic. For definiteness, we take the primitive operators of the latter to be negation (¬) and conjunction (&), with other truth-functional operators being treated as metalinguistic abbreviations, for example material implication (⊃). A *frame* for this language is simply an ordered pair <W, R>, where W is a nonempty set and R a binary relation over W. Since R is informally understood as the accessibility relation for a knowledge operator, which is factive, we require R to be reflexive. By itself, a frame does not interpret the language, since it does not assign propositions to its formulas, in particular to its atomic formulas. To do that, we need an *interpretation* (or *model*) over the frame. An interpretation over <W, R> is a function I mapping each formula *A* of the language to a subset I(*A*) of W (a proposition), satisfying three recursive conditions to ensure that it interprets the sentential operators as intended:

$$\mathrm{I}(\neg A) = \mathrm{W} - \mathrm{I}(A)$$

$$\mathrm{I}(A \,\&\, B) = \mathrm{I}(A) \cap \mathrm{I}(B)$$

$$\mathrm{I}(KA) = \mathrm{KI}(A) = \{w \in \mathrm{W}\text{: } \forall x (\text{if } \mathrm{R}wx,\ x \in \mathrm{I}(A))\}$$

Here *K* operates on sentences while K operates on propositions, as defined in section 3. A formula *A* is *valid* over a frame <W, R> if and only if for every interpretation I over <W, R>, I(*A*) = W; in other words, *A* is true at every world on every interpretation. Thus <W, R> invalidates *A* if and only if I(*A*) ≠ W for some interpretation I over <W, R>.

The aim is to show that if a formula is invalid on some epistemic frame, it is invalid on some epistemic frame satisfying the constraints (1a)-(1e). The first step is to provide a way of combining a simple unmarked clock frame <Θ, Δ> with an arbitrary epistemic frame <W, R> to make a new epistemic frame <W×Θ, R×Δ>. Informally, a new world in W×Θ is simply a world in W together with a position for the clock's hand. One new world is accessible from another just in case both component worlds of the former are accessible in the relevant sense from the respective component worlds of the latter: to discriminate between new worlds, the observer must discriminate either between their first components or between their second components. About <W, R> we assume only that R is reflexive, as above.

We first describe the simple clock frame <Θ, Δ> more carefully and check that it satisfies (1a)-(1e). For simplicity, we take Θ to be the set of points on a circle in standard Euclidean space, though it can easily be adjusted to allow Θ to be countable or finite. Since the worlds in Θ are just positions, we set [θ] = θ and *r** = *r*. Thus (1a) holds trivially in <Θ, Δ>. We measure the distance |η, θ| between positions η and θ by the relevant angle in degrees; thus 0 ≤ |η, θ| ≤ 180. We fix a constant *c* such that 0 < *c* < 180 and define the indiscriminability relation Δ to hold between η and θ just in case |η, θ| ≤ *c*. Thus Δ is reflexive, holds between some but not all pairs of positions on the circle, and obviously satisfies the monotonicity condition (1b). It also satisfies the invariance condition (1c) since |*r*(η), *r*(θ)| = |η, θ| for any rotation *r*. Now let *z* be any position and θ the position diametrically opposite *z*, so |*z*, θ| = 180, so Δ*z*θ fails (since *c* < 180). Moreover, P_θ = {θ}, so *z* ∈ K(W−P_θ), so <Θ, Δ> satisfies (1d). Moreover, if *s* is a rotation through *c* degrees, *s* ≠ $\underline{1}$ (since 0 < *c*) and |*z*, *s*(*z*)| = *c*, so Δ*zs*(*z*), so (1e) holds too. Thus the simple clock frame <Θ, Δ> is reflexive and satisfies all the constraints (1a)-(1e).

We now construct the new frame <W×Θ, R×Δ>. We define:

W×Θ = {<*w*, θ>: *w* ∈ W and θ ∈ Θ}

R×Δ = {<<w_1, $\theta_{1>,<}$$w_2$, θ_2>>: <w_1, w_2> ∈ R and <θ_1, θ_2> ∈ Δ}

Thus the new worlds are just ordered pairs of old worlds and positions on the circle; the new accessibility relation just requires the old accessibility relations to hold with respect to each component. The function which specifies the hand's position at one of the new worlds is defined simply by setting [<*w*, θ>] = θ; correspondingly, the proposition P_θ = {<*w*, θ>: *w* ∈ W}. To rotate new worlds through an angle of θ we merely rotate the component representing their position on the circle:

*r**(<*w*, θ>) = <*w*, *r*(θ)>

Since any position on the circle can be rotated to any other, the orbit of a world <*w*, θ> is simply {<*w*, η>: η ∈ Θ}. The distance measure on a given orbit simply follows that on the positions:

|<*w*, η>, <*w*, θ>| = |η, θ|

Given these definitions, we can easily check that <W×Θ, R×Δ> satisfies constraints (1a)-(1c), because <Θ, Δ> already does. Now let <*w*, *z*> be any new world. We first check (1d) for <W×Θ, R×Δ>. In checking it for <Θ, Δ>, we already established that Δ*z*θ fails for some position θ. Thus (R×Δ)<*w*, *z*><*x*, θ> also fails for any *x* in W, so W−P_θ is true at any new world to which <w, z> has R×Δ, so K(W−P_θ) is true at <w, z>, as (1d) requires. Now we check (1d) for <W×Θ, R×Δ>. In checking it for <Θ, Δ>, we showed $\Delta zs(z)$ for some rotation $s \neq \underline{1}$; since R is reflexive by hypothesis, (R×Δ)<w, z><w, $s(z)$>, as (1e) requires. Moreover, since both R and Δ are reflexive, so is R×Δ. Thus the new frame <W×Θ, R×Δ> is reflexive and satisfies the constraints (1a)-(1e).

The next task is to show that if a formula is invalid on <W, R>, it is also invalid on <W×Θ, R×Δ>. By contrast, the converse is not generally true. For example, the formula $Kp \supset KKp$ (where p is an atomic formula), which expresses the KK principle, is invalid on <W×Θ, R×Δ>, by what has already been proved. But we can choose R to be a transitive relation, making the formula valid on <W, R>.

First, we define a 'translation' of propositions in <W, R> into propositions in <W×Θ, R×Δ>. Informally, the idea is simply that W×Θ divides each way in which a proposition in W can be true into many, which differ from each other only in the position of the hand. Thus for X ⊆ W, we define tr(X) = {<w, θ>: w ∈ X, θ ∈ Θ}. Hence the truth-value of tr(X) remains constant as we vary just the second component of the world in W×Θ. We need the 'translation' tr to respect the propositional operations expressed in the language, corresponding to negation, conjunction, and the knowledge operator K, in this sense, for all subsets X and Y of W:

$$\mathrm{tr}(W-X) = (W\times\Theta)-\mathrm{tr}(X)$$

$$\mathrm{tr}(X \cap Y) = \mathrm{tr}(X) \cap \mathrm{tr}(Y)$$

$$\mathrm{tr}(KX) = K\mathrm{tr}(X)$$

In the third equation, note that strictly speaking 'K' expresses an operation on subsets of W on the left-hand side but an operation on subsets of W×Θ on the right-hand side; in practice, context always resolves the ambiguity. The reader is spared the routine checks that the three equations hold.

Let I be any interpretation of the formal language over the old frame <W, R>. We define a corresponding mapping I^{tr} from formulas to subsets of W×Θ by setting $I^{tr}(A)$ = tr(I(A)). We check that I^{tr} is an interpretation of the language over the new frame <W×Θ, R×Δ>, in other words, it respects the sentential operators over <W×Θ, R×Δ>, given that I respects them over <W, R>. That is easy:

$$I^{tr}(\neg A) = \mathrm{tr}(I(\neg A)) = \mathrm{tr}(W-I(A)) = (W\times\Theta)-\mathrm{tr}(I(A)) = (W\times\Theta)-I^{tr}(A)$$

$$I^{tr}(A \ \&\ B) = \mathrm{tr}(I(A \ \&\ B)) = \mathrm{tr}(I(A) \cap I(B)) = \mathrm{tr}(I(A)) \cap \mathrm{tr}(I(B)) = I^{tr}(A) \cap I^{tr}(B))$$

$$I^{tr}(KA) = \mathrm{tr}(I(KA)) = \mathrm{tr}(KI(A)) = K\mathrm{tr}(I(A)) = KI^{tr}(A)$$

Thus I^{tr} is an interpretation of the language over <W×Θ, R×Δ>.

We can now complete the proof of the desired result. Suppose that an epistemic frame <W, R> invalidates a formula A. Thus for some interpretation I over <W, R>, I(A) ≠ W.

But for any subset X of W, tr(X) = W×Θ only if X = W. Hence $I^{tr}(A)$ = tr(I(*A*)) ≠ W×Θ. Since I^{tr} is an interpretation over <W×Θ, R×Δ>, that frame invalidates *A* too. Thus any formula invalid on some reflexive frame is invalid on some reflexive frame satisfying the constraints (1a)-(1e). QED. To put the result in more positive terms: any formula satisfiable on an epistemic frame is satisfiable on an epistemic frame satisfying the constraints.

The result shows that (1a)-(1e) commit one to no new principles in the propositional language beyond those of the most elementary epistemic logic. An example is the symmetry of the accessibility relation (as opposed to rotational symmetry!). The accessibility relation for the clock frame <Θ, Δ> is symmetric, for |η, θ| = |θ, η|. This corresponds to the validity of the formula *p* ⊃ *K*¬*K*¬*p* on <Θ, Δ>. However, we can still choose a frame <W, R> with R non-symmetric, so that *p* ⊃ *K*¬*K*¬*p* is invalid on <W, R>. Then *p* ⊃ *K*¬*K*¬*p* is also invalid on <W×Θ, R×Δ>. For example, if R*wx* holds without R*xw*, then (R×Δ)<*w*, θ><*x*, θ> holds without (R×Δ)<*x*, θ><*w*, θ> for any θ in Θ, so R×Δ is non-symmetric too. All invalidities but not all validities are preserved from the old frames to the new. Thus the constraints (1a)-(1e) do *not* force special rules in this language holding across the simple clock models to be treated as general principles about knowledge.

Of course, (1a)-(1e) themselves can be formalized in a more expressive language, without being completely general principles about knowledge—actuality may contain no such clock. If we simply introduce quantifiers into sentence position, modelled as ranging over all sets of worlds, we can express the existence (not just possibility) of exceptions to the KK principle in the formula ∃*p* (*Kp* & ¬*KKp*). It will be valid in every combined frame <W×Θ, R×Δ>, even when R itself is transitive, since at each world <*w*, θ> a verifying value of the variable *p* is {<*x*, η>: R*wx*, Δθη}, for that assignment automatically makes *Kp* true at <*w*, θ>, and it makes *KKp* false at <*w*, θ>, for the structure of the simple clock model guarantees that for some positions η and ζ, Δθη and Δηζ but not Δθζ, so (R×Δ)<*w*, θ><*w*, η> and (R×Δ)<*w*, η><*w*, ζ> (since R is reflexive) but not (R×Δ)<*w*, θ><*w*, ζ>, so *p* is false at <*w*, ζ>, so *Kp* is false at <*w*, η>, so *KKp* is false at <*w*, θ>. Thus the formula saying that there are counterexamples to KK is valid in combined frames.

5. *Transitive rotation-invariant monotonic relations in clock frames*

From section 3, we can extract a more general argument applicable to other relations which epistemologists might invoke to understand clock cases. Such relations should be *rotation-invariant*, invariant under rotational automorphisms of the frame, in the same sense as R. For a binary relation Q over W, this is just the analogue of (1c), for any rotational automorphism *r** corresponding to a rotation *r* of the circle, and any worlds *w* and *x*:

(2c) Q*r**(*w*)*r**(*x*) if and only if Q*wx*

Subject to the same glosses as for (1c), it would be arbitrary to single out a relation Q violating (2c) for any special theoretical privilege: why prefer Q to its image under rotation by *r**?

We also restrict consideration to relations which satisfy the same monotonicity constraint as R; (2b) is analogous to (1b):

(2b) If w, x, and y are in the same orbit, and $|w, x| \geq |w, y|$, then Qwx only if Qwy

Unlike rotation-invariance, monotonicity is not a requirement of non-arbitrariness. Again unlike rotation-invariance, monotonicity is also not preserved under negation: (2c) entails the result of substituting ¬Q for Q in (2c), but (2b) does not entail the result of substituting ¬Q for Q in (2b). Instead, (2b) holds when Q is some sort of *similarity* relation of epistemological significance. For if w, x, and y are in the same orbit, and $|w, x| \geq |w, y|$, then y is at least as similar to w as x is in epistemologically significant respects, since y differs from x at most in having been rotated less far from w.

We do *not* assume that Q is the accessibility relation for an operator for some familiar epistemic matter such as knowledge or justified belief. Q may be embedded deeper inside the workings of an epistemological theory than that.

We *do* assume that the circle of positions is in standard Euclidean space, with no infinitesimal distances.

How does a *transitive*, rotation-invariant, monotonic relation behave on a single orbit? There are few possibilities. For suppose Qwx, where w and x are distinct worlds in the same orbit, so $|w, x| > 0$. Using rotation-invariance (2c), monotonicity (2b), and an argument like that in section 3, we can show that w is joined to any world in its orbit by a finite chain of worlds, each of which has Q to the next. Hence, if Q is transitive, w has Q to every world in its orbit. Therefore, by rotation-invariance, *every* world in the orbit has Q to every world in the orbit: on this orbit, Q coincides with the universal relation. An alternative is that Q *never* holds between distinct worlds on this orbit. Then if Q holds between some world and itself, by rotation-invariance Q holds between *every* world and itself on the same orbit: on this orbit, Q coincides with identity. The only remaining alternative is that Q coincides with the empty relation. Thus, on any given orbit, a transitive, rotation-invariant, monotonic relation is either universal or empty or identity. Conversely, one can easily check that the universal relation, the empty relation, and identity are all transitive, rotation-invariant, and monotonic. Thus, on a given orbit, there are exactly three possibilities for the extension of a transitive, rotation-invariant, monotonic relation.

If Q is supposed to be some sort of similarity relation, it is presumably reflexive. That eliminates the possibility of its coinciding with the empty relation on an orbit. Consequently, a transitive, rotation-invariant, monotonic similarity relation on a given orbit holds either universally or just in case of identity. Only those two extreme possibilities remain. That result is relevant to the final section.

All three constraints—transitivity, rotation-invariance, and monotonicity—are needed to restrict the possibilities so drastically: dropping any one of them opens up many more possibilities. We can see this by checking each pair of constraints on any given orbit, as follows.

Rotation-invariance and monotonicity without transitivity: For any real number c in the open interval (0, 180), define: Q$_c wx$ just in case $|w, x| < c$. Then Q$_c$ is rotation-invariant and monotonic but not transitive. Whenever $c \neq d$, $Q_c \neq Q_d$, so there are uncountably many such relations.

Transitivity and rotation-invariance without monotonicity: For any real number c in the open interval (0, 180), define: $Q^c wx$ just in case $|w, x| = nc$ for some natural number n. Then Q^c is transitive and rotation-invariant but not monotonic. Whenever the ratio c/d is irrational, $Q^c \neq Q^d$, so there are uncountably many such relations.

Transitivity and monotonicity without rotation-invariance: For any real number c in the half-open interval (0, 180] and world w in the given orbit, define: $Q_{c,w}xy$ just in case either $x = w$ and $|w, y| \leq c$ or $x = y$. Then $Q_{c,w}$ is transitive and monotonic but not rotation-invariant. Whenever $c \neq d$ or $v \neq w$, $Q_{c,v} \neq Q_{d,w}$, so there are uncountably many such relations.

If these examples are representative, dropping transitivity yields far more natural results than dropping either rotation-invariance or monotonicity.

Rotation-invariance is just one example of symmetry in a frame. Symmetry is an abstract phenomenon; in itself, it does not exert pressure on epistemology in any particular direction. Nevertheless, we have seen, as a constraint it can drastically simplify the options, by clearing out a mess of *ad hoc* alternatives. It would be interesting to find other applications of such symmetry constraints in epistemology, or indeed elsewhere in philosophy.

6. *Comparisons*

Clock frames make a good test for defences of the KK principle. The present aim is not to provide detailed critiques of such defences, but rather to highlight a few key features.

One approach to knowledge is to understand it in terms of the flow of information (Dretske 1981, Greco 2014: 180-7, Stalnaker 2019: 38-48). Observing the clock puts the observer into a state which carries information about the state of the clock, and in particular about the position of the hand. Information flows from the clock to the observer when conditions are ‘normal’ (Greco) or the ‘channel conditions’ obtain (Stalnaker). We use Greco’s more familiar terminology. For a state S of the observer, let S^c be the most specific state of the clock such that, necessarily, if the observer is in S and conditions are normal, the clock is in S^c. When conditions *are* normal and the observer *is* in S, state S carries the information that the clock is in S^c. That is supposed to put the observer in a position to know that the clock is in S^c, and whatever follows from the clock’s being in S^c. To take advantage of the opportunity and actually gain the knowledge, the observer must actually form the corresponding belief, in a way suitably dependent on being in S, but the main cognitive challenge is seen as already met by the flow of information from the clock to the observer.

In what follows, for simplicity, knowing and being in a position to know are conflated, but distinguishing them will not undermine the main line of argument. The reader is at liberty to substitute ‘is in a position to know’ for ‘knows’ in what follows.

How does such an approach support the KK principle? The idea is that when conditions are normal and, as observer, one is in state S, that state carries not just the information that the clock is in state S^c but also the further information that one *knows* that the clock is in S^c. The argument goes thus. By hypothesis, necessarily, if one is in S and conditions are normal, the clock is in S^c. The kind of necessity at issue can legitimately be assumed to be maximally broad and so to satisfy the principle that necessities are necessarily necessary; the normality condition can still be used to restrict possible worlds to those of current interest. Thus it is necessarily necessary that if one is in S and conditions are normal, the clock is in S^c. Hence, *a fortiori*, necessarily, if one is in S and conditions are normal, then necessarily, if one is in S and conditions are normal, the clock is in S^c. So if one is in S and conditions are normal, the state S carries the information that one knows that the clock is in S^c. Thus one knows that one knows that the clock is in S^c. When conditions are *ab*normal, on the simplest version of the view, one knows nothing, so the KK principle holds vacuously.

A problem for such arguments is that what constitutes normality is itself contingent and changeable. In the intended sense, it is now normal for women to vote, but it was not normal two centuries ago. Contingency and changeability infect even normal conditions for vision: normality in lighting, air clarity, and visual discrimination are contingent and changeable too. If the clause 'conditions are normal' applies to all worlds (or situations) which are normal by their own standards, it will apply far too widely for epistemological purposes: the observer's states will carry too little information, because the intended connections break down in worlds utterly unlike our own but normal by their own standards. If instead 'conditions are normal' applies only to worlds which are normal by *our* standards, it is inappropriate for determining what information a state carries in a world where normality differs from actual normality.

The natural alternative is to treat 'conditions are normal' as expressing an accessibility relation which a world *w* has to a world *x* just in case *x* is normal by the standards of *w*. We can use the word 'normally' to express the necessity operator restricted by that accessibility relation. Thus the key move in defence of KK was from 'Normally, if one is in S, the clock is in S^c' to 'Normally, if one is in S, then normally, if one is in S, the clock is in S^c'. But the inference pattern from 'Normally, if *A*, *C*' to 'Normally, if *A*, then normally, if *A*, *C*' is truth-preserving only if the inference pattern from 'Normally *C*' to 'Normally normally *C*' is truth-preserving (consider the special case where *A* is a tautology and so redundant). In effect, the argument goes from iteration for normality to iteration for knowledge. But iteration is implausible for normality (see Carter 2019 and Loets 2019 for detailed discussion). For example, think of a time when it had very recently become normal for women to vote. Someone may then have found it refreshing that it was normal for women to vote, precisely because it was not yet normal for it to be normal for women to vote. One can construct modal analogues of such examples. A similar problem arises for articulating the restriction with the sentence 'Channel conditions obtain' instead: the deep trouble is not interference from the ordinary meaning of the word 'normal' but the postulated role of the restriction to the favoured worlds in a theory of knowledge.

For present purposes, the information-carrying account is most easily applied to the clock case if the normality or channel conditions are held fixed for the sake of argument. An obvious consequence of the account is that when one is in state S, one's state trivially carries the information that one is in S, so one knows that one is in S, irrespective of how specific S. How does this play out in clock frames?

To implement the normality approach, we can add to the frame both a set N (⊆ W) of normal worlds and a function S mapping each world *w* to the observer's state S(*w*) in *w*. In the spirit of clock frames, N will be rotation-invariant: no position of the hand is more normal than any other. The same-state relation between worlds (S(*w*) = S(*x*)) will also be rotation-invariant, since the function S has no intrinsic bias against positions of the hand . It is automatically transitive, and as a similarity relation it is also monotonic. Thus identity of one's state is a transitive, rotation-invariant, monotonic relation between worlds. Therefore, on a given orbit, it coincides either with the universal relation or with identity, by what was noted in section 5. If it coincides with the universal relation, then one's state is the same irrespective of the hand's position, so on the information-carrying account one learns nothing about the position by looking at the clock. That is just scepticism about the external world, not at all what friends of the account want. Consequently, they must go for the only alternative: on a given orbit, the same-state relation coincides with identity. Thus, however finely the hand's positions are individuated, one's states must be equally finely individuated. Since each state carries the information that one is in it, however finely the states are individuated, one's state carries the information that one is in exactly that state. What makes that more than a trivial definitional matter is the epistemological significance accorded it by the information-carrying account.

Of course, to know that one is in state S, one still has to form the belief that one is in S. That is where all the cognitive work needs to be done. But that makes the information-carrying story something of a sideshow. Imagine a perfectly reliable mechanism which forms a perfectly accurate image of the clock-face in one's brain. Thus one is in a state which carries exact information about the position of the hand, but one still faces the task of discriminating the position of the hand in the image, of *reading* the internal clock to answer the question 'What time does it show?' The theorist can posit a homunculus in the brain whose job is to observe the image and come to know the position of the hand in it, but there is an obvious threat of a regress. To take a more realistic example, that one's total state trivially carries exact information about one's blood pressure does not trivialize the process of measuring one's own blood pressure. By itself, the mere internality of one's states has no epistemological significance.

The proper challenge is to understand human knowledge *without* helping oneself to an unexplained level of perfect self-transparency. Since transparency is usually associated with explicit attributions of it to conscious states by internalist epistemologies, it is easily overlooked when it is implicitly attributed to physical states by an externalist epistemology, but that does not make it more legitimate.

The same point emerges when we consider epistemic frames for the information-carrying account. The relevant aspects of a world *w* are the exact state O(*w*) of the observer and the specific state C(*w*) of the clock. The natural information-theoretic definition of the accessibility relation is this: R*wx* just in case (i) O(*w*) = O(*x*) (the observer is in the same state) and (ii) if *w* is normal, so is *x*. If *w* is a normal world, C(*w*) is a determinate of O(*w*)^c (the unspecific clock state normally corresponding to the observer state O(*w*)). For any other determinate C of $O(w)^c$, $O(x) = O(w)$ and $C(x) = C$ for some normal world *x* (so R*wx* and R*xw*). The proposition that the clock is in $O(w)^c$ is just {*w*: C(*w*) is a determinate of $O(w)^c$}. This set-up has the intended effect that, in a normal world *w*, one knows that the clock is in $O(w)^c$, and knows nothing stronger purely about the clock. Such frames validate self-transparency in the sense that in any world *w*, one knows that one is in O(*w).*

KK is valid in such frames, as intended: R is transitive, so all instances of the positive introspection principle *Kp* ⊃ *KKp* are true at all worlds on all interpretations. Much more controversial is the *negative* introspection principle ¬*Kp* ⊃ *K*¬*Kp*: if you don't know, you know that you don't know. Negative introspection is invalid in such frames. For let *w* be a normal world and *x* a world such that O(*x*) = O(*w*) but C(*x*) is not a determinate of $O(x)^c$, so *x* is abnormal. Thus *w* is accessible from *x* but *x* is inaccessible from *w*. Negative introspection fails in *x*, for one doesn't know that things are normal (since they are not normal), but one also doesn't know that one doesn't know that things are normal (since for all one knows one is in *w*, where things are normal). However, all instances of negative introspection are true at all *normal* worlds on all interpretations. For only normal worlds are accessible from normal worlds, and on normal worlds accessibility is equivalent to identity of the observer's state, which is an equivalence relation, and therefore validates negative introspection. For example, suppose that slightly rotating the normal world *w* clockwise yields the equally normal world *y*, where C(*w*) is within the $O(w)^c$ region but near its edge going anti-clockwise, so C(*w*) excludes $O(y)^c$, although C(*w*) and C(*y*) are determinates of $O(w)^c$ and $O(y)^c$ respectively. Thus in *w* the clock is not in $O(y)^c$, so one doesn't know that the clock is in $O(y)^c$. Suppose R*wz*. Then R*zw* because accessibility is symmetric on normal worlds, so in *z* one doesn't know that the clock is in $O(y)^c$ because in the accessible world *w* the clock is not in $O(y)^c$. Thus in no world accessible from *w* does one know that the clock is in $O(y)^c$, so in *w* one knows that one doesn't know that the clock is in $O(y)^c$. This is wonderfully fine-grained knowledge of one's epistemic state, for *w* is very close to *y*, in which one knows, and so knows that one knows, that the clock *is* in $O(y)^c$. In all normal worlds, one knows exactly what one's knowledge state is. That is an even higher level of self-knowledge than satisfaction of KK.

The qualified commitment to negative introspection can be reduced by *grading* normality. Thus condition (ii) in the definition of normality, that *x* is normal if *w* is, can be replaced by the graded condition that *x* is at least as normal as *w*. Since the normality ordering is transitive, this modification preserves positive introspection, but more normal worlds will be accessible from corresponding less normal ones and not *vice versa*, so non-maximally normal worlds will fail negative introspection. But negative introspection will still hold at maximally normal worlds (for the given condition O(*w*)). Thus, for all one knows even in less normal states, one satisfies negative introspection with respect to all propositions. But that is not always so. After all, one can *know* on the basis of reliable testimony that one violates negative introspection (the testifier does not specify with respect to which propositions).

An alternative to the information-carrying strategy is to emphasize the epistemic challenge of extracting knowledge from information-carrying states. But if that is where much of the epistemological action takes place, the question is whether one can extract knowledge that one knows that P from an information-carrying state from which one can extract knowledge that P—which seems depressingly close to the original question.

Das and Salow (2018) focus much more on the final step of forming the belief that one knows. Very roughly, they argue that when one knows X, one can use that very knowledge as a knowledge-constituting basis for the belief that one knows X (via a quasi-inferential step from 'X' to 'I know X'), and thereby come to know that one knows X. The details of their argument are too intricate to rehearse here. However, one key aspect deserves mention.

Das and Salow work with the popular framework in which the epistemic status of a belief depends on its *basis*. For example, in determining the reliability of a particular belief, the relevant comparison class is the set of beliefs with the same basis. The basis is content-specific, not something generic, such as *vision*. For example, they propose that, on a safety-theoretic version of their view, one should gloss 'safe' along these lines: 'an agent's belief that *p*, formed on basis *B*, is safe if that agent couldn't easily have falsely believed that *p* on that basis' (2018: 10). Clearly, *B* must be closely tailored to *p* for that gloss to work properly. As they note, the effect of such a gloss is very sensitive to the individuation of bases. Much of their discussion concerns the same-basis relation, and its application to beliefs that one knows formed by their proposed rule. At least to a first approximation, the relevant comparison class for a given belief b_0 in a world w_0 contains a belief *b* in a world *w* if and only if the basis of b_0 in w_0 = the basis of *b* in *w*. Thus sameness of basis is an equivalence relation between something like belief-world pairs: it is reflexive, symmetric, and transitive. Moreover, its grain should be neither maximally fine nor maximally coarse. If it were super-fine, it would be too kind to bad true beliefs, by excluding nearby false beliefs from their comparison classes. If it were super-coarse, it would be too unkind to good true beliefs, by including distant false beliefs in their comparison classes.

Sameness of basis is a similarity relation in the sense of section 5. Therefore, to deserve a starring role in epistemological theory, it should be both monotonic and rotation-invariant in clock cases. But it is also transitive, as an immediate consequence of its definition. Therefore, by the argument of section 5, in a clock case it should coincide on a given orbit either with the universal relation or with identity. But the former option is too coarse and the latter too fine for the epistemological role of bases. The result of a rotation through 1° should be in the relevant comparison class; the result of a rotation through 180° should not be. But rotation-invariance and transitivity together are incompatible with that combination.

The natural solution is to move from identity to similarity in how the belief is formed. Such similarity is non-transitive. Oddly, when Das and Salow *motivate* the role of bases in epistemological theory, they write in terms of graded similarity. For example, in the sentence preceding the one quoted above, they mention 'a false belief formed *on a very different basis*' (their emphasis); '*very*' is relevant only to a graded similarity relation. In a footnote to the latter sentence, they mention Keith DeRose's requirement 'that possibilities resemble the actual world in a particular way to count as ones that "could easily have happened" in the relevant sense'. They add: 'We're inclined to think that talk of bases is mostly a way of putting a label on the resemblance in question' (2018: 20n18). Again, they say 'the basis of pre-theoretically similar beliefs should be individuated similarly' (2018: 11);

why expect pre-theoretic similarity to be transitive? Those informal motivating remarks do nothing to favour a transitive similarity relation over a non-transitive one. But they are used to motivate official formulations in their arguments that turn on identity of basis. For example, in surveying the main theoretical options for individuating the basis of a belief formed by following a rule, they consider only standards for identity of basis (2018: 15). In effect, transitivity has slipped in between the motivation and what it is supposed to motivate. Given the close connection between transitivity and the KK principle, that is no small lacuna in an argument for the latter.

How might reformulating the issues in terms of non-transitive *basic similarity* rather than *identity of basis* affect Das and Salow's arguments? Their proposal is that if one knows X, one can use *knowing X* as a basis on which to believe that one knows X; since a belief on that basis that one knows X is guaranteed true, one thereby knows that one knows X. What is that proposal analogous to in terms of non-transitive basic similarity?

The key epistemological role of the basis of a belief is to define a comparison class of other (possible) beliefs with which to assess the original belief. Substituting basic similarity for identity of basis does not obviate the need for a comparison class. For simplicity, we may treat the comparison class as a class of worlds rather than beliefs, and hold the content of the belief fixed, as in standard epistemic logic. Nothing crucial in the argument will depend on that simplification.

One can define a transitive similarity relation R^+ in terms of a non-transitive similarity relation R by stipulation: R^+wx if and only if for all z, Rwz just in case Rxz. However, R^+ is typically too restrictive to demarcate a useful comparison class. In the clock case, if R holds between positions on the circle just in case the angular distance between them is n^o ($0 < n < 180^o$), R^+ is simply identity: the only comparison is self-comparison.

Instead, the natural proposal is that the comparison class for a world is simply the class of basically similar worlds, so that basic similarity plays the role of a non-transitive accessibility relation in standard epistemic logic. Then one may know X in a world *w* because X is true in every world in the comparison class for *w*. But, even though the comparison class is held constant, it by no means follows that one *knows* X in every world in the comparison class for *w*, since the comparison class for a world *x* in the comparison class for *w* may include a world *y* not in the comparison class for *w* (that is just to restate non-transitivity). Thus, without transitivity, there is no proper analogue of the idea that whenever one knows, one's knowing is available as a basis for belief. Such talk is not a harmless rephrasing of talk about basic similarity; it is a Trojan horse for transitivity, and so for the KK principle.

By the argument of section 5, in the case of the unmarked clock, the kind of transitive same-basis relation Das and Salow postulate will be either non-monotonic or non-translation-invariant. Either it will flout the underlying similarity measure or it will be *ad hoc*.

As these reflections show, the model-building approach facilitates the identification of key structural problems for epistemological theories. That is best done by focus not on particular models, but on broad classes of model, defined by structural constraints. One important type of structural constraint is invariance under a given group of symmetries of the epistemic frame, which helps eliminate clutter and *ad hoc* fixes. By using these and other structural constraints, we can make the results of the model-building approach more robust.[2]

Notes

[1] Proof: (i) When accessibility is transitive: Suppose that one knows a proposition (set of worlds) X in a world *w*. Let *x* be any world accessible from *w*, and *y* any world accessible from *x*. By transitivity, *y* is accessible from *w*, so X is true in *y*. Thus X is true in any world accessible from *x*, so in *x* one knows X. Thus one knows X in any world accessible from *w*, so in *w* one knows that one knows X. Thus in any world in which one knows X, one knows that one knows X. (ii) When accessibility is not transitive: Let *w*, *x*, and *y* be worlds such that *x* is accessible from *w*, *y* is accessible from *x*, but *y* is not accessible from *w*. Let X be the proposition true in all and only the worlds accessible from *w*. Thus in *w* one knows X. Since *y* is not accessible from *w*, X is false in *y*. Since *y* is accessible from *x*, in *x* one does not know X. Since *x* is accessible from *w*, in *w* one does not know that one knows X. Thus the KK principle fails at *w*.

[2] Versions of this paper were presented to the 2019 Analytic Philosophy Symposium at the University of Texas, Austin, the 2020 Workshop on Luck, Risk, and Competence at the University of Seville, and a class at Oxford. I thank the audiences for helpful discussion, and Sam Carter, Kevin Dorst, and Bernhard Salow for detailed written comments.

References

Carter, Sam. 2019: ‘Higher order ignorance inside the margins’, *Philosophical Studies*, 176: 1789-1806.

Das, Nilanjan, and Salow, Bernhard. 2018: ‘Transparency and the KK principle’, *Noûs*, 52: 3 -23.

Dorst, Kevin. 2019: ‘Abominable KK failures’, *Mind*, 128: 1227-1259.

Dretske, Fred. 1981: *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Goodman, Jeremy, and Salow, Bernhard. 2018: ‘Taking a chance on KK’, *Philosophical Studies*, 175: 183-196.

Greco, Daniel. 2014: ‘Could KK be OK?’, *The Journal of Philosophy*, 111: 169-197.

Loets, Annina. 2019: ‘Choice points for a theory of normality’, MS.

McHugh, Conor. 2010: ‘Self-knowledge and the KK principle’, *Synthese*, 173: 231-257.

Stalnaker, Robert. 2015: ‘Luminosity and the KK Thesis’, in Sanford Goldberg (ed.), *Externalism, Self-knowledge, and Scepticism*: 17-40. Cambridge: Cambridge University Press. Reprinted in Stalnaker 2019.

Stalnaker, Robert. 2019: *Knowledge and Conditionals: Essays on the Structure of Inquiry*. Oxford: Oxford University Press.

Williamson, Timothy. 1992: ‘Inexact knowledge’, *Mind*, 101: 217-241.

Williamson, Timothy. 2000: *Knowledge and its Limits*. Oxford: Oxford University Press.

Williamson, Timothy. 2007: *The Philosophy of Philosophy*. Oxford: Blackwell.

Williamson, Timothy. 2011: ‘Improbable knowing’, in Trent Dougherty (ed.), *Evidentialism and its Discontents*: 147-164. Oxford: Oxford University Press.

Williamson, Timothy. 2013: ‘Gettier cases in epistemic logic’, *Inquiry*, 56: 1-14.

Williamson, Timothy. 2014: ‘Very improbable knowing’, *Erkenntnis*, 79: 971-999.

Williamson, Timothy. 2017: ‘Model-building in philosophy’, in Russell Blackford and Damien Broderick (eds.), *Philosophy’s Future: The Problem of Philosophical Progress*: 159-173. Oxford: Blackwell-Wiley.