EXPLICATING A STANDARD EXTERNALIST ARGUMENT AGAINST THE KK PRINCIPLE

Simon D'ALFONSO

ABSTRACT: The KK principle is typically rejected in externalist accounts of knowledge. However, a standard general argument for this rejection is in need of a supportive explication. In a recent paper, Samir Okasha argues that the standard externalist argument in question is fallacious. In this paper I start off with some critical discussion of Okasha's analysis before suggesting an alternative way in which an externalist might successfully present such a case. I then further explore this issue via a look at how Fred Dretske's externalist epistemology, one of the exemplifying accounts, can explain failure of the KK principle.

KEYWORDS: knowledge, externalism, reliabilism, KK principle, Fred Dretske

According to various versions of the KK (Knowing that One Knows) principle, if one knows some proposition *p* then they know or are at least in a position to know that they know that *p*. This principle is often taken to be a dividing factor between internalist and externalist accounts of knowledge, with the former typically endorsing the principle and the latter typically rejecting it. A general explanation of one of the standard externalist arguments for rejecting the KK principle, exemplified by the reliabilist family of knowledge accounts, is given in the following passage:

> For, if warrant may be external to our cognitive perspective, then there is no special reason to expect those who know that p to be in a position to know that their belief that p is warranted. This can be seen this more clearly by focusing on the reliabilist theory of knowledge. If one's belief that p is produced by a reliable process that one knows nothing about, then one may have no way of knowing that this belief constitutes knowledge, and thus no way of knowing that one knows that p.[1]

Thus according to this externalist position, an epistemic agent can know that *p* without knowing that their belief that *p* was reliably produced (the externalist condition). But if knowledge is defined as reliably produced true belief, then knowing that they know that *p* implies knowing that their belief that *p* was

[1] David Hemp, "The KK (knowing that one knows) principle," *Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/kk-princ/#H1, last accessed 6 September 2013.

reliably produced. As suggested by Samir Okasha[2] though, relying on this conflict to reject the KK principle proves to be problematic and the standard argument against the KK principle that uses it is fallacious.

Okasha represents the generic externalist line of reasoning in question with the following formal reasoning in epistemic logic:

(1) Kp	(assumption)
(2) $\neg K[Bpisreliable]$	(assumption)
(3) $Kp\equiv[Bp\wedge p\wedge Bpisreliable]$	(externalism definition)
(4) $Kp\Rightarrow KKp$	(KK principle, assumed for *reductio*)
(5) KKp	(from (1), (4))
(6) $K[Bp\wedge p\wedge Bpisreliable]$	(from (3), (5))
(7) $K[Bpisreliable]$	(from (6), assuming knowledge distributes across conjunction)
(8) $\neg[Kp\Rightarrow KKp]$	(from (1), (2), (3), (4), (7), by *RAA*)

As he points out though, this argument is fallacious, as the derivation of (6) from (3) and (5) involves substitution within an intensional context; specifically, substituting the externalist definition of knowledge for the second occurrence of ‘K’ in the expression ‘KKp.’ Might there be a way however of deriving (6) and rejecting the KK principle that does not involve this problematic move?

Appealing to the closure of knowledge under known implication ($Kp\wedge K(p\Rightarrow q)\vdash Kq$) offers one plausible way to derive (6), provided that a further assumption is made regarding knowledge of the definition of knowledge. The reasoning then becomes:

(1) Kp	(assumption)
(2) $\neg K[Bpisreliable]$	(assumption)
(3) $K(Kp\equiv[Bp\wedge p\wedge Bpisreliable])$	(knowledge of externalist definition)
(4) $KKp\Rightarrow K[Bp\wedge p\wedge Bpisreliable]$	(from distribution on (3))
(5) $Kp\Rightarrow KKp$	(KK principle, assumed for *reductio*)
(6) $K[Bp\wedge p\wedge Bpisreliable]$	((from (1), (4), (5)))
(7) $K[Bpisreliable]$	(from (6), assuming knowledge distributes across conjunction)
(8) $\neg[Kp\Rightarrow KKp]$	(from (2), (7), by *RAA*)

Despite this possibility, Okasha claims that

> this weaker principle is insufficient to get us from (3) and (5) to (6); since the equivalence $Kp\equiv[Bp\wedge p\wedge Bpisreliable]$, and thus the implication

[2] Samir Okasha, “On a flawed argument against the KK principle,” *Analysis* 73 (2013): 80–86.

> $Kp\Rightarrow[Bp\wedge p\wedge Bpisreliable]$, is presumably not something that the typical subject knows or believes. At most, a handful of reliabilist epistemologists might know that this implication holds.[3]

Whilst this might be the case, it must be stressed that we generally take a rejection of the KK principle here to mean that it fails to hold in certain cases. This position is not anti-KK; that is, it is not committed to denying that there can be cases where KKp follows from Kp. It thus suffices to show that there can be certain cases within this argument setting where Kp holds and KKp does not. Now, even if $K(K\equiv[Bp\wedge p\wedge Bpisreliable])$ is only true in cases where the K is the knowledge operator of an externalist epistemologist who knows the definition of knowledge, this suffices to show that there are cases where the KK principle does not hold. The result is that the KK principle would fail in cases where the knowledge operator is that of an externalist epistemologist who knows the externalist definition of knowledge!

In fact, knowledge of the definition is not required. We could start off with the weaker $KKp\equiv K[Bp\wedge p\wedge Bpisreliable]$. If we dissect this equivalence further, we come up with the following three implications:

(1) $KKp\Rightarrow KBp$

(2) $KKp\Rightarrow Kp$

(3) $KKp\Rightarrow K[Bpisreliable]$

(1) and (2) are uncontroversial axioms in an epistemic-doxastic logic. It all really boils down to formula (3). If (3) is accepted then the KK principle would fail to hold as per the reasoning outlined above.

But even if $K(K\equiv[Bp\wedge p\wedge Bpisreliable])$ were to be the case, it could be claimed that "externalists typically reject even this weaker closure principle ... [so] ... it seems unlikely that a commitment to closure is behind the inference from (3) and (5) to (6)."[4] This point however is problematically cursory and demonstrates a failure to be careful in distinguishing between a rejection of the universality of a principle and an outright rejection of the principle. Rejecting closure does not mean denying that such an inference ever holds. Rather, it means that in certain circumstances and for certain reasons, Kp and $K(p\Rightarrow q)$ can be true whilst Kq is false, for some propositions p and q. Going by the externalist accounts I am familiar with, there is no reason to suggest that they would deny closure in such a case.

[3] Okasha, "On a flawed argument," 83.

[4] Okasha, "On a flawed argument," 83.

Beyond the analysis thus far, might we be able to present this externalist point against the KK principle in a different way, one that is not susceptible to the problem above? I think that there is; in fact, it is how I first analysed this externalist position on the KK principle.

In the argument above, the premise $\neg K[Bpisreliable]$ is used to form the final contradiction. This premise however should be replaced with another, one that forms part of an iterative application of the externalist definition of knowledge, with *Kp* being the proposition that is known. According to this application, *KKp* if and only if:

1. BKp

2. Kp

3. $B[Kp]isreliable$

Thus the KK principle question can be put as

Kp⇒(BKp∧Kp∧B[Kp]isreliable)?

Of these three conditions listed above, we take the first two to hold given *Kp*. Condition 2 (*Kp*) trivially follows from the consequent. Condition 1 (*BKp*) has been used as an argument against the KK principle, for if the KK principle is valid then Kp implies that BKp and thus knowledge would be ruled out for agents that are not capable of introspection or forming second-order beliefs (Kelp *et al* 2011).[5] But let us set that aside and confine ourselves to knowledge and agents for which $Bp\Rightarrow BBp$ and $Kp\Rightarrow BKp$ holds. This leaves us with condition 3 (B[Kp] is reliable), which can fail to follow from *Kp*. The externalist line of reasoning thus becomes:

(1) Kp	(assumption)
(2) $\neg(B[Kp]isreliable)$	(assumption)
(3) $Kp\equiv[Bp\wedge p\wedge Bpisreliable]$	(externalism definition)
(4) $KKp\equiv[BKp\wedge Kp\wedge B[Kp]isreliable]$	(substituting *Kp* for *p* in (3))
(5) $Kp\Rightarrow KKp$	(assumed KK principle)
(6) KKp	(from (1) and (5))
(7) $BKp\wedge Kp\wedge B[Kp]isreliable$	(from (4) and (6))
(8) $B[Kp]isreliable$	(conjunct of (7))
(9) $\neg[Kp\Rightarrow KKp]$	(*RAA* from (2) and (8))

[5] Christoph Kelp and Nikolaj J.L.L. Pedersen, "Second-Order Knowledge," in *The Routledge Companion to Epistemology*, ed. Sven Bernecker and Duncan Pritchard (London: Routledge, 2011), 586–596.

So in order for one to know that they know that *p* in this externalist framework, their belief that they know that *p* must be reliable. But this does not necessarily follow from *Kp* and so the KK principle fails to hold.

Now, this discussion has all been in terms of a generic notion of externalist reliability. In actuality the array of externalist accounts are going to have their own specific externalist correlate and how exactly it can be the case that *Kp* whilst ¬(*B*[*Kp*]*isreliable*), is something for a given externalist account to explain. In order to exemplify this externalist line of reasoning and also for the sake of generally discussing the issue, I shall now attempt to fit this analysis around Dretske's prominent externalist account of knowledge.

One of the pioneers of externalism and relevant alternative accounts in epistemology,[6] Dretske gave a complete embodiment of these ideas via his informational account of knowledge.[7] In short, according to this account knowledge is defined as information-caused belief. Information carrying and belief causing signals provide the externalist component and a veridical conception of information ensures that such beliefs are true. For Dretske:

> A signal *r* carries the information that *s* is *F* = The conditional probability of *s*'s being *F*, given *r* (and *k*), is 1 (but, given *k* alone, less than 1)

A set of relevant alternatives is used to determine the range of possibilities over which this probability assessment is made. Dretske's classic zebra example[8] provides an amusing and simple way to relate these ideas. In this scenario, one is at a zoo where they go to the zebra section and see what they believe to be a zebra. Now, we can say that every time one receives a 'zebra' visual signal in certain circumstances then that means there is a zebra before them and so the required probability is 1. Therefore this visual signal carries the information that there is a zebra and any resulting belief that there is a zebra is knowledge. But for this judgement on the information-carrying status of the signal to be made, certain theoretically possible but non-actual irrelevant alternatives, such as ones in which the creature before the zoo goer is a cleverly disguised mule painted to resemble a zebra or a virtual zebra in a simulation, are not considered. Thus it is within the set of relevant alternatives, which say, consist of standard zoo scenarios, that a 'zebra' visual signal carries the information that there is a zebra.

[6] See Fred Dretske, "Conclusive Reasons," *Australasian Journal of Philosophy* 49 (1971): 1–22, Tim Black, "Contextualism in Epistemology," *Internet Encyclopedia of Philosophy*, 2006, http://www.iep.utm.edu/contextu/#SH3a, last accessed 6 September 2013.

[7] See Fred Dretske, *Knowledge and the Flow of Information* (Cambridge: MIT Press, 1981).

[8] See Fred Dretske, "Epistemic Operators," *Journal of Philosophy* 67 (1970): 1007–1023.

This all means that one can see a zebra at the zoo and know that there is a zebra before them without being able to ascertain that the creature is not a painted mule or a virtual zebra simulation. Such irrelevant alternatives determine a set of conditions: the signal carries the information that 'zebra' given that 'painted mule' is not the case, 'virtual zebra' is not the case, etc. That these conditions obtain is not something the knower has to know though and as the following passage from Dretske captures, it is this that can lead to failure of the KK principle:

> modest contextualism (and, hence, externalism) provides an illuminating explanation of why KK fails. It fails because factual knowledge, according to modest contextualism, depends for its existence on circumstances of which the knower may be entirely ignorant. So the knower can know that P without knowing (as required by KK) that he knows that P.[9]

So the zebra example is a candidate for a situation where the KK principle fails; whilst one knows that there is a zebra before them (Kz) they need not be in a position to know this (KKz). How could this work exactly?

With the strategy I have in mind, the gist involved here is that whilst one can get the required information and know some proposition within the bounds of certain relevant alternatives, the inability to be certain within a wider context and rule out certain irrelevant alternatives precludes their second-order knowledge of that proposition.

In judging $Kz \Rightarrow KKz$, we treat Kz as a proposition that falls within the scope of the outer K operator. Given this, KKz would require the satisfaction of the following three conditions:

1. *Kz*

2. *BKz*

3. The belief that *Kz* being caused by the information that *Kz* (reliability equivalent).

Conditions 1 and 2 are straightforward enough and were covered in discussion above. Condition 3 is the interesting condition and the one that fails under this externalist framework. In order to explain why it fails here, we need to explain how the information that *z* is present whilst the information that *Kz* is not. Such an explanation is available given that the set of relevant alternatives against which the information that *Kz* is judged is different to and greater than the set of relevant alternatives against which the information that *z* is judged. This is

[9] Fred Dretske, "Externalism and modest contextualism," *Erkenntnis* 61 (2004): 176.

related to the other infamous feature of externalist epistemologies such as Dretske's, namely the failure of knowledge under known implication. One can know that something is a zebra and know that something being a zebra implies that it is not a mule ($\neg m$) without knowing that it is not a disguised mule. This is because the set of relevant alternatives for the proposition $\neg m$ includes 'disguised mule', something that cannot be ruled out with a visual zebra signal.

For a belief that Kz to be caused by the information that Kz would require a signal that carries sufficient information coupled with an agent's reliable mechanism for forming beliefs about their beliefs as knowledge. If we take it as a given that the latter will be consistent across all of the relevant alternatives, then it is the insufficiency of the visual zebra signal that leads to a lack of meta-knowledge. This insufficiency is due to the fact that as with the case of closure and $\neg m$, the proposition Kz is judged against a more demanding set of relevant alternatives than z, a set in which possibilities such as disguised mule are included. Since the signal does not carry the information that z relative to this more demanding set of relevant alternatives and does not carry the information that it carries the information that z, then it is not enough for Kz either. In this way, we can also say that the visual zebra signal does not suffice to carry the information that Kz because there are relevant scenarios (like the painted mule one) where the 'zebra' visual signal occurs and Bz results but $\neg z$ and therefore $\neg Kz$.

In light of this discussion, it is worth briefly noting that we might roughly portray the logical connection between closure and the KK principle with the following arguments:

- $Kp \wedge (\exists Q)(K(p \Rightarrow Q) \wedge \neg KQ) \vdash \neg KKp$
- $KKp \vdash \neg(\exists Q)(K(p \Rightarrow Q) \wedge \neg KQ)$

with another argument to consider being what I term KK closure:

- $\mathrm{KKp} \wedge \mathrm{K}(\mathrm{p} \Rightarrow \mathrm{q}) \vdash \mathrm{Kq}$

A point to be made from all of this is that the approach taken towards relevant alternatives will determine whether and how these principles hold. If one is an attributive[10] or 'radical' contextualist as Dretske puts it,[11] then they posit an interpretation of the relevant alternatives idea whereby each of the premises involved in the evaluation of a knowledge/information argument shares the same set of relevant alternatives. In such a case, the KK principle and closure under

[10] Patrick Rysiew, "Epistemic Contextualism," *The Stanford Encyclopedia of Philosophy*, 2011, http://plato.stanford.edu/archives/win2011/entries/contextualism-epistemology/, last accessed 6 September 2013.

[11] Dretske, "Externalism and modest contextualism."

known implication will be valid. It is for relevant alternatives theorists such as Dretske, for whom the set of alternatives can differ between propositions,[12] that these principles are rejected.

It is also the case for Dretske that the information carried by a signal for an agent is in part determined by what they already know and what relevant alternatives they are in an epistemic position to rule out. With regards to the formalised arguments above, it is here that we can also see an important connection between $\neg K[Bpisreliable]$ and $\neg B[Kp]isreliable$. If the zoo goer in the zebra-mule scenario does not know that their belief that *z* is reliable in the sense that they are not in a position to rule out irrelevant alternatives such as 'mule', then their belief that *Kz* will not be reliable in the sense that it is not caused by the information that Kz. Thus we might posit the following principle: $\neg K[Bpisreliable] \Rightarrow \neg B[Kp]isreliable$.

As long as the standards for *Kp* are higher than the standards for *p*, the KK principle can fail. But if one comes to know *p* by meeting the standards of *Kp*, then unless some reason other than the externalist one covered here intervenes both *Kp* and *KKp* will be true. For example, in the standard zebra-mule scenario suppose that the relevant alternatives for *z* consist of all standard zoo scenarios and the relevant alternatives for *Kz* consist of all scenarios in which any animal is in the enclosure. Within these parameters whilst a 'zebra' visual signal does not suffice for *KKz*, if the method used to determine the type of animal was a DNA test instead, then the information carried by this result would be enough for both *Kz* and *KKz*.

Thus this particular externalist idea of KK principle failure relies on a variation of relevant alternative sets for the first-level proposition *p* and the second-level proposition *Kp*. In certain applications both sets could be limited to alternatives, such as those in the zebra example in the previous paragraph, that have a distinguishing set of information and can in practice or principle ultimately be uniquely determined. In this way *KKp* could be met. On the other hand, if the relevant alternatives for *Kp* include extreme skeptical alternatives such as brain-in-a-vat scenarios, then *KKp* will never be met (unless we find a way to obtain information that rules out such scenarios!).

[12] Steven Luper, "The Epistemic Closure Principle," *Stanford Encyclopedia of Philosophy*, 2010, http://plato.stanford.edu/entries/closure-epistemic/#CloFaiRelAltApp, last accessed 6 September 2013.