NORMALITY, SAFETY AND KNOWLEDGE

Markos Valaris

UNSW Sydney

Recent epistemology has seen a striking rise in interest in the notion of *normality*, including in the analysis of justified belief, defeasible reasoning, and knowledge. In the analysis of knowledge in particular, normality has been used to support *modal* analyses of knowledge, according to which knowledge is safely true belief. In this paper, I sound a note of caution regarding this proposal. As I will argue, the counterexamples that originally seemed to threaten the safety analysis of knowledge in its more traditional formulations have natural counterparts that continue to threaten the newer, normality-based formulations. Moreover, these reformulated counterexamples seem to exploit structural features of the notion of normality itself, rather than one or another particular conception of normality.

1. Introduction

Recent epistemology has seen a striking increase interest in the notion of *normality*, including in the analysis of justified belief (Smith 2010; 2016; Goodman and Salow 2018), defeasible reasoning (McHugh and Way 2016; Valaris 2017), and knowledge (Greco 2016; Dutant 2016; Goodman and Salow 2018; Beddor and Pavese 2020). This seems like a promising development. Nonetheless, the aim of the present paper is to sound a note of caution. As of now, we do not have an agreed-upon and fully fleshed-out account of normality. In the absence of such an account, it is reasonable to worry that at least some of the claims made on behalf of normality in epistemology may prove to be over-ambitious. I will suggest that this may, indeed, be the case, when it comes to one particular project: the analysis of knowledge.

The analysis of knowledge has, of course, been a long-standing ambition of epistemology. Rather notoriously, no proposal has gained universal acceptance. Still, one of the most popular approaches has been a modal one, motivated by the thought that knowledge requires not just true belief, but belief that is non-accidentally true. This idea is often developed in terms of *safety*. Informally, for an agent to know it is not enough for her to believe truly; it must also be the case that she would not *easily* have believed falsely. More officially:

> *Safety*: *S* knows that *p* if and only if in all (or nearly all) relevant worlds in which *S* has the same (or similar[1]) belief on the same basis, the belief is true.[2]

The reasons for favouring a condition like *Safety* are well-known, and I will not rehearse them here. Our concern is with how exactly to specify the class of worlds in which you need to avoid believing falsely.

Early discussions of safety relied on a relation of *similarity* to actuality to identify the relevant worlds: what can easily happen, on such a view, is what does happen in the worlds most similar to actuality.[3] Such views, however, face counterexamples, in both directions: there can be seemingly unsafe beliefs that intuitively count as knowledge, as well as seemingly safe beliefs that do not. This is where normality comes in as a potential solution: if we specify the worlds relevant to assessing the application of *Safety* in terms of *normality* rather than similarity then, it has been

---

[1] This condition is needed to handle beliefs in necessary truths, which would otherwise come out as trivially safe. This issue will play no role in what follows.

[2] In this formulation, safety is taken to be both necessary *and* sufficient for knowledge, as required by the analytical project. This view is endorsed by Pritchard (2005) (who, however, has since changed his mind), Lasonen-Aarnio (2010), and Beddor and Pavese (2020). Others take it to be a necessary condition only (Sosa 1999; Williamson 2000; Pritchard 2012). Still others develop formal models in which knowledge is represented as safe belief (e.g., Williamson 2013) without necessarily committing to such an analysis. The focus here is on the prospects of safety as an *analysis* of knowledge, in the traditional sense.

[3] See, e.g., Sainsbury (1997), Sosa (1999), Williamson (2000) and Pritchard (2005). This way of understanding modal closeness is familiar from classic accounts of counterfactuals (Stalnaker 1968; Lewis 1973).

argued, the counterexamples may be avoided.[4] On such a view, worlds in which you go wrong and which are very similar to actuality do not count as relevant and can be ignored, provided they are sufficiently *abnormal*; conversely, worlds in which you go wrong and which are quite different from actuality may threaten the safety of your beliefs, provided that they are not. Despite its initial promise, however, this view faces trouble of its own. Indeed, as I will argue, counterexamples similar to those that afflicted the similarity-based accounts appear to plague the normality-based ones too.

Before we move on, one final methodological remark. Proponents of normality-based epistemology have tended not to commit to a particular substantive account of normality. This makes it hard to assess how exactly their proposals fare against counterexamples. In what follows, I will not hold them to any particular account. I will, however, assume that they are committed to offering an informative analysis of knowledge, and hence that they are committed to there being *some* principled way of evaluating worlds or states of affairs for normality, which broadly corresponds with pre-theoretical intuitions about how that latter notion works. This relatively weak demand, as we will see, proves surprisingly difficult to satisfy.

2. Safety, Similarity, Normality

To get the discussion going, let us review a pair of examples that appear to speak against *Safety*. Here is one such example, adapted from Neta and Rohrbaugh (2004, 400). This example targets the necessity direction of *Safety*:

Lab

[4] The clearest example of such an argument is found in Beddor and Pavese (2020). Dutant (2016) also suggests understanding safety in terms of normality as a way to avoid the counterexamples, but he does not aim for an analysis of knowledge in the traditional sense. Greco (2016), Goodman and Salow (2018), Dutant and Littlejohn (forthcoming), and Carter and Goldstein (2021) also link knowledge to normality, but their motivations are different.

> In preparation for a psychological experiment, you are given a glass of orange juice. You happen to have been assigned to the control group, so your glass contains plain orange juice. Had you been assigned to the treatment group, it would have been mixed with a drug that would have impaired your memory. As a member of the control group, you are shown seven flashes. Further, you *know* you have been shown seven flashes. Had you been assigned to the treatment group, however, you would have been shown only six flashes, but—due to the effects of the drug—you would still have believed you had been shown seven.

Cases with this structure at least *appear* to be cases of knowledge without safety.[5] For one thing, your subjective experience need not be different in any way in a world in which you have been assigned to the treatment group. Nor need there be any other major differences from actuality in such a world. Thus, it seems, there are worlds very similar to actuality in which you go wrong in your memory-based beliefs. Your actual memory-based belief, therefore, would appear to be unsafe. Since in the case described you *know* you have been shown seven flashes, *Safety* appears to give the wrong answer.

Here is an apparent counterexample to the sufficiency of *Safety*, from Pritchard (2012, 260)[6]:

> *Temp*
>
> Temp forms his beliefs about the temperature in the room by consulting a thermometer. His beliefs, so formed, are highly reliable, in that any belief he forms on this basis will always be correct. Moreover, he has no reason for thinking that there is anything amiss with his thermometer. But the thermometer is in fact broken and is fluctuating randomly within a given range. Unbeknownst to Temp, there is an agent hidden in the room who

[5] For related counterexamples, see Kelp (2009) and Bogardus (2014).

[6] Roush (2005, chap. 2) considers a structurally similar counterexample.

> is in control of the thermostat [and] whose job it is to ensure that every time Temp consults the thermometer the "reading" on the thermometer corresponds to the temperature in the room.

The intuition here is that Temp does not know the temperature in the room. After all, his thermometer is broken, and the readings turn out to be correct only through the intervention of another agent, whose very existence is unknown to Temp. And yet, given the setup, Temp's beliefs about the temperature appear to satisfy *Safety*: it would take a world significantly *dissimilar to* actuality (a world without the helper) to make Temp go wrong.

How should proponents of *Safety* respond? I focus on one particular style of response, suggested by Dutant (2016) and Beddor and Pavese (2020): replacing the similarity-based criterion in *Safety* with one based on normality. More specifically, Beddor and Pavese (2020, 69-71) suggest that the worlds in which you need to avoid falsity are those which are *at least as normal as the actual world for the performance of the relevant task*. What does normality, relative to a task, amount to? Beddor and Pavese do not provide an explicit account, but suggest that we rely on our intuitions regarding the conditions "we would consider *fair* for the performance and assessment of the task" (2020, 69).

Here is how the appeal to normality is meant to help. In the *Lab* case, the worlds in which you mistakenly believe that you were shown seven flashes are ones where you have ingested a drug that messes with your memory. In an intuitive sense, then, such worlds seem abnormal. But then, it seems, believing falsely in *those* worlds (even if they are the ones most similar to actuality) should not count against your claim to knowledge in the actual world, where your memory is not messed with in this way. Setting those worlds aside, however, there is no reason to doubt that your belief that you were shown seven flashes is safe.

Conversely, in the Temp example abnormal circumstances are responsible for the agent's success in the *actual* case: it is only because of the intervention of the helper that Temp's beliefs about the room's temperature end up being true. Accordingly, it is to worlds where such

abnormality is *absent* that we should look if we are to fairly assess his epistemic performance—even if those worlds involve a greater departure from actuality. Given that Temp's thermometer is broken, however, in many of those worlds Temp ends up with false beliefs about the temperature; hence, his actual belief, although true, is not safe.

Problem solved, then? Unfortunately, things are not so simple. There are straightforward ways to generate new counterexamples, similar to the ones just discussed, but which are not handled by the appeal to normality.

3. The Counterexamples Revisited

Let us begin, as before, with the necessity direction of *Safety*. Consider the following case:

> *Enhancement*
>
> In preparation for a psychological experiment, you are given a glass of orange juice. As it happens, you have been assigned to the treatment group for an experimental drug that enhances your capacities for perceptual discrimination, and especially subitizing. Subsequently, you are briefly shown a pattern of dots and asked to report on their number. You confidently report (and know) that there are seven dots in the pattern. Had you been assigned to the control group, you would have been shown a pattern with six dots—but you would still have reported (and believed) that the pattern contained seven.

The example, of course, mirrors *Lab* from the previous section. It exploits the following fact: just as there can be interventions that abnormally *degrade* our cognitive capacities, there can also be interventions that abnormally *enhance* them. But while normality appears to be symmetric in this way (both enhancements and impediments may count as abnormal), epistemic assessments don't seem to be: your performance while in the control group for the enhancement experiment counts as knowledge, despite the existence of worlds in which you go wrong and which are, if anything, *more* normal for the performance of the relevant task.

How can proponents of *Safety* respond? Perhaps one might be tempted to deny that, in a case like this, you really have knowledge. I think this is very hard to accept. It is hard to deny that temporarily enhanced cognitive abilities can provide us with knowledge that would be out of reach otherwise. Regardless of what one thinks of the particular example presented above, so long as this is granted the trouble for the view is clear.

Could the example be avoided by adopting sufficiently fine-grained descriptions of the relevant tasks (as Beddor and Pavese (2020, 69) at one point suggest)? Clearly, if we were to distinguish between the tasks that you perform in the treatment and control conditions in *Enhancement*, the case would no longer present a problem. But I do not think this is a promising way to proceed. For one thing, there is an obvious and salient sense in which the task is the *same* in the two conditions: that's the point of doing controlled experiments in psychology and other sciences in the first place. Furthermore, if we are allowed to use fine-grained task descriptions as we please, we lose the earlier diagnosis in *Temp*: Temp's performance in the task of *telling the temperature with the helper's assistance* is, after all, quite safe.

The argument so far has relied on an intuitive understanding of normality relative to a task. But it also seems to be vindicated on at least some independently motivated conceptions of normality available in the literature.

One influential approach to normality has been developed by Martin Smith (2010; 2016). The basic idea is that, as he puts it, "normal conditions require less *explanation* than abnormal ones do" (2016, 39).[7] The intuition here is that the normality of a state of affairs corresponds to how well that state of affairs can be accounted for by the explanatory regularities that prevail in a given world. A state of affairs that *cannot* be accounted for in this way will need its own, *ad hoc* explanation. Thus, whether a world *w* is more or less normal than actuality depends on how well *w* conforms to what would be predicted by the explanatory regularities that hold in actuality.

[7] For a related view, see also Pietroski and Rey's (1995) account of *ceteris paribus* laws.

Notably, while no world can be more similar to actuality than actuality is to itself, there is no barrier to worlds *more normal* than actuality. This is because many of the explanatory regularities that hold in actuality do not hold exceptionlessly, but only *ceteris paribus*; thus, it will in general be possible to find worlds that match these regularities better than actuality does (Smith 2016, 111-113).

Now, Smith's central aim is to model the relation of evidential support, not the notion of conditions normal for the performance of a task. But we can adapt his discussion to our context in the following way:

> Conditions *C* count as *less normal* than conditions *C** for the performance of a task *T* just in case your performance (success or failure) in *C* is *less readily explained* by a relevant set of explanatory regularities compared to *C**.

But then, it seems that the treatment condition is *more* abnormal than the control. In the control condition, your failure is explained by the limits of the human capacity for subitizing; it is your success in the treatment condition that requires a special explanation.

On a rather different approach to normality, normality is closely related to *typicality* (Dutant 2016; Hawthorne and Carter (ms)). On this approach, a sequence of ten coin-tosses that all come up tails will be counted as *more abnormal* than, say, one whose outcome looks like this: HHTHTTTHHT. This verdict is inconsistent with Smith's understanding of normality, since of course neither outcome requires any more explanation than the other (Smith 2017). Rather, on this approach the reason why getting ten tails in a row counts as abnormal has to do with the fact that it exhibits a highly improbable *pattern* (or higher-order property): among the 1024 equiprobable outcomes of our experiment, it is the only one that contains no heads. But, again, this approach does not seem to help: your enhanced ability to subitize in the treatment condition is surely less typical than its ordinary human-level counterpart.

Is there some other way to understand normality, so as to avoid the counterexample? As we saw earlier, Beddor and Pavese (2020,69) link normality to the conditions we would deem "fair" for task performance and evaluation. Could this perhaps help?[8]

The trouble with this suggestion is that, if the appeal to fairness is to give the right results in *Enhancement*, the link between normality and fairness will have to be sacrificed. On any intuitive understanding of normality, the conditions under which you perform in *Enhancement* are *abnormally* good. If we nonetheless insist on counting them as fair, we must be using a notion of fairness that allows for (some) abnormal conditions to be counted as fair. Perhaps we do indeed make use of some such notion of fairness in our epistemic thinking: after all, we seem prepared to credit you with knowledge (an epistemic achievement) in cases like *Enhancement*. If, however, this is how fairness is to be understood, it cannot help with the project of analysing knowledge in terms of normality.

Let us now turn to the sufficiency direction. Problems arise here as well. We can bring this out by adding a twist to the *Temp* example discussed in the previous section:

> *Incurious Temp*
>
> Just as Temp, except Temp normally has no interest at all in the temperature of the room. In fact, he can only be induced to look at the thermometer (which, as before, is broken) by the influence of his helper, who happens to possess psychic powers.

With this modification, the response outlined earlier no longer appears to work. Worlds without a helper are not worlds in which Temp has false beliefs about the temperature, but worlds in which he has no beliefs about the temperature at all. To find worlds in which Temp goes wrong we need to look at worlds in which the helper causes Temp to look at his thermometer but subsequently fails to adjust the temperature to match the reading. Given the description of the case, however, such worlds appear to be less normal than actuality: after all, helping Temp is

[8] I thank an anonymous referee for urging me to consider this suggestion.

what the helper *normally* does. Counter-intuitively, therefore, *Safety* appears to predict that Temp knows the temperature of the room.

While *Enhancement* exploits the fact that enhancements can be as abnormal as impediments, *Incurious Temp* exploits the fact that assessments of normality are subject to trade-offs. A state of affairs may be more abnormal than another in one respect, but that abnormality may be offset by its being less abnormal in others. This is what happens in our example. The helper's assistance renders actuality more abnormal, in one respect, than a case in which Temp arrives at his beliefs about the temperature unaided. Given our stipulations, however, worlds in which Temp arrives at his beliefs about the temperature unaided need to be abnormal in other, offsetting respects. In the *most* normal worlds Temp simply forms no beliefs about the temperature at all.[9]

Once again, the verdict appears to be vindicated by independently motivated accounts of normality. Consider Smith's (2010; 2016) explanatory account first. The worlds in which Temp forms false beliefs about the temperature require more explanation than the worlds in which Temp forms true ones, since in those worlds Temp goes wrong *despite* the helper's presence: perhaps the thermostat the helper uses to adjust the temperature is broken, or perhaps she changes her mind at the last minute, or… No matter how we fill in the details, such worlds appear to require more *ad hoc* explanatory posits than actuality.

[9] An anonymous referee suggests that, among those most normal worlds from which the helper is absent, there will be some in which Temp just happens to glance at the thermometer, thereby forming (false) beliefs about the temperature in the room. The example could be amended to close this loophole (suppose that Temp, being incurious about the temperature, does not even own a thermometer; it is only in worlds with the helper that the thermometer is present in the room at all). More important, however, is to be clear about the structural point the example illustrates. There is nothing to prevent cases in which an agent's forming false beliefs about a certain subject matter would be more abnormal than her forming true beliefs about it, even though the *way* in which she forms her true beliefs remains, intuitively, of the wrong type to provide knowledge.

It is harder to know what to say about this case on an account that links normality with typicality. But if we think of typicality in terms of the probability of patterns, it seems that worlds in which Temp goes wrong about the temperature are, again, more abnormal than ones in which he gets it right. By hypothesis, worlds in which Temp goes wrong about the temperature are worlds in which the helper is present. Such worlds may well be less typical than worlds without a helper at all. But among the worlds *with* a helper, the more typical ones will be ones in which Temp gets it *right*, not ones in which he gets it wrong. This suffices to secure the verdict that the latter worlds are more abnormal than the former.

As it turns out, therefore, the appeal to normality does not insulate *Safety* from counterexamples. Furthermore, the counterexamples do not hinge on idiosyncratic features of any particular substantive theory of normality; rather, they exploit structural features that would appear to be shared by any theory that stays true to our intuitive notion. I cannot, of course, show that no future normality-based safety account will be able to handle these problems. At present, however, I do not see how.[10]

Beddor, B. & Pavese C. (2020). Modal Virtue Epistemology. *Philosophy and Phenomenological Research* 101 (1): 61–79. Doi: doi.org/10.1111/phpr.12562.

Bogardus, T. (2014). Knowledge Under Threat. *Philosophy and Phenomenological Research* 88 (2): 289–313. Doi: doi.org/10.1111/j.1933-1592.2011.00564.x.

Carter, S. & Goldstein S. (2021). The Normality of Error. *Philosophical Studies* 178 (8): 2509–33. Doi: doi.org/10.1007/s11098-020-01560-6.

Dutant, J. (2016). How to Be an Infallibilist. *Philosophical Issues* 26 (1): 148–71. Doi: doi.org/10.1111/phis.12085.

Goodman, J. & Salow B. (2018). Taking a Chance on KK. *Philosophical Studies* 175 (1): 183–96. Doi: doi.org/10.1007/s11098-017-0861-1.

Greco, D. (2016). Safety, Explanation, Iteration. *Philosophical Issues* 26 (1): 187–208. Doi: doi.org/10.1111/phis.12067.

Hawthorne, J. & Carter S. Normality (ms). Accessed 4 January 2022. htpps://www.academia.edu/50107734/Normality.

[10] I wish to thank two anonymous referees for this journal for their thoughtful and constructive comments. Research for this paper was supported by the Australian Research Council (grant number DP200101045).

Kelp, C. (2009). Knowledge and Safety. *Journal of Philosophical Research* 34: 21–31. Doi: doi.org/10.5840/jpr_2009_1.
Lasonen-Aarnio, M. (2010). Unreasonable Knowledge. *Philosophical Perspectives* 24 (1): 1–21. Doi: doi.org/10.1111/j.1520-8583.2010.00183.x.
Lewis, D. (1973). *Counterfactuals*. Malden, MA: Blackwell Publishers.
Littlejohn, C. & Dutant J. forthcoming. Justification, Knowledge, and Normality. *Philosophical Studies*, 1–17. Doi: doi.org/10.1007/s11098-019-01276-2.
McHugh, C. & Way J. (2016). What Is Good Reasoning? *Philosophy and Phenomenological Research* 92 (3). Doi: doi.org/10.1111/phpr.12299.
Neta, R. & Rohrbaugh, G. (2004). Luminosity and the Safety of Knowledge. *Pacific Philosophical Quarterly* 85 (4): 396–406. Doi: doi.org/10.1111/j.1468-0114.2004.00207.x.
Pietroski, P. & Rey G. (1995). When Other Things Aren't Equal: Saving Ceteris Paribus Laws from Vacuity. *The British Journal for the Philosophy of Science* 46 (1): 81–110. Doi: doi.org/10.1093/bjps/46.1.81
Pritchard, D. (2005). *Epistemic Luck*. Oxford: Oxford University Press.
———. (2012). Anti-Luck Virtue Epistemology. *Journal of Philosophy* 109 (3): 247–79. Doi: doi.org/10.5840/jphil201210939.
Roush, S. (2005). *Tracking Truth: Knowledge, Evidence, and Science*. Oxford University Press.
Sainsbury, M. (1997). Easy Possibilities. *Philosophy and Phenomenological Research* 57 (4): 907–19. Doi: doi.org/10.2307/2953809
Smith, M. (2010). What Else Justification Could Be. *Noûs* 44 (1): 10–31. Doi: doi.org/10.1111/j.1468-0068.2009.00729.x.
———. 2016. *Between Probability and Certainty: What Justifies Belief*. Oxford: Oxford University Press.
———. (2017). Why Throwing 92 Heads in a Row Is Not Surprising. *Philosopher's Imprint* 17 (21). http://hdl.handle.net/2027/spo.3521354.0017.021.
Sosa, E. (1999). How to Defeat Opposition to Moore. *Philosophical Perspectives* 13: 137–49. Doi: doi.org/10.1111/0029-4624.33.s13.7.
Stalnaker, R. (1968). A Theory of Conditionals. In N Rescher (ed.) *Studies in Logical Theory* (98–112). Malden, MA: Blackwell.
Valaris, M. (2017). Induction, Normality and Reasoning with Arbitrary Objects. *Ratio* 30 (2): 137–48. Doi: doi.org/10.1111/rati.12129.
Williamson, T. (2000). *Knowledge and Its Limits*. New York: Oxford University Press.
———. (2013). Gettier Cases in Epistemic Logic. *Inquiry* 56 (1): 1–14. Doi: doi.org/10.1080/0020174X.2013.775010.