Counterfactuals and the Epistemology of Modality

Thomas Kroedel
Humboldt University of Berlin

1. The Explanatory Project

We seem to have plenty of modal knowledge, that is, knowledge that such-and-such is metaphysically necessary and knowledge that so-and-so is metaphysically possible. For instance, we know that necessarily everything that is coloured is spatially extended, and we know that there might have been flying pigs. Modal knowledge has been particularly attractive to philosophers. For they generally do not confine themselves to claiming that such-and-such happens to be the case but might have been otherwise; rather, they tend to make the stronger claim that such-and-such has to be the case. As Nozick once put it: “necessity lures philosophers. It is the flame, the philosopher the moth” (2001: 120). To the extent that philosophers disagree with one another, they also tend to make claims about possibility, since in order to refute that such-and-such is necessary, one has to show that the negation of such-and-such is possible after all. Thus, if there is such a thing as philosophical knowledge, a large proportion of it will be modal knowledge. All the same, philosophers have, until very recently, been fairly nonchalant about how we know truths about necessity and possibility. How can this knowledge be explained? This paper tries to answer that question.

What does it mean to explain a given kind of knowledge? We could give a direct explanation if there were a set of non-epistemic conditions that were strictly sufficient for knowledge; knowledge could then be explained by showing that these conditions are satisfied, and modal knowledge in particular could be explained by showing that our corresponding modal beliefs satisfy these conditions. However, the repeated failure of epistemologists to analyse knowledge ever since the discovery of Gettier cases strongly suggests that no strictly


2. The requirement that the conditions be non-epistemic is supposed to rule out circularity, as in the ‘explanation’ of someone’s knowing p by her seeing p, remembering p, etc.
sufficient non-epistemic conditions for knowledge are available.\textsuperscript{3} I shall therefore take an indirect approach to an explanation of modal knowledge. Instead of starting from strictly sufficient conditions for knowledge, I suggest explaining modal knowledge by explaining the reliability of the processes that form our corresponding modal beliefs — that is, by explaining the tendency of these processes to form true beliefs. Epistemic reliability is not strictly sufficient for knowledge: I may reliably, and truly, believe that my lottery ticket will lose and yet not know that my lottery ticket will lose.\textsuperscript{4} Still, reliability has a good claim to be at least a strictly necessary condition for knowledge. More importantly, reliability captures a salient feature of knowledge, viz., the idea that someone who knows could not easily have erred. Accounting for this salient feature of modal knowledge would get us a long way towards explaining the latter. This, at any rate, is what I shall assume for the remainder of the paper. (Those not convinced by the approach of explaining knowledge from reliability might still read it as an attempt to account for the reliability of the processes that form our modal beliefs, which is a worthy epistemological project in itself.)

Generally, evolution may explain why certain belief-forming processes should be reliable: if having true beliefs of a certain kind increases an individual's chances of survival and reproduction, then, \textit{ceteris paribus}, processes that tend to produce true beliefs of this kind will be selected for.\textsuperscript{5} In order to apply an explanation of this kind to the processes that form our modal beliefs, we would have to show that it is useful, or fitness-enhancing, to be reliable in modal matters. On the face of it, having mainly true modal beliefs is not a terribly useful thing. It may be conducive to survival in a population of academic philosophers, but out in the wilderness it hardly seems to be of any help. It does not follow, however, that no evolutionary explanation of the reliability of the processes that form our modal beliefs is forthcoming. Only a simplistic view of evolution would claim that a trait cannot be explained by evolution unless it is useful itself. This view overlooks the fact that a trait might also be explained by being a byproduct of some useful trait.\textsuperscript{6} Often, feature $F_1$ is useful while feature $F_2$ is rather useless, so that having features $F_1$ and $F_2$ is no more useful than having $F_1$ alone, but $F_1$ and $F_2$ together are much easier to implement than $F_1$ without $F_2$. For instance, the human ability to dance does not seem to have much survival value in itself, while the general ability for coordinated movement in climbing, fighting, etc. has. But it seems that an ability for coordinated movement in climbing, fighting, etc. that did not bring with it the ability to dance would be comparatively difficult to implement. Thus, we could explain the ability to dance as an evolutionary byproduct of more useful abilities for coordinated movement. Analogously, we might explain our ability to make correct modal judgements as the byproduct of some more useful ability that is implemented more easily with the former than without it. What could such a useful ability be? In order to answer this question, we need to know more about the processes that form our modal beliefs. We need to know what the mechanisms of modal thought look like in more detail and how they are related to other cognitive abilities.\textsuperscript{7}

Accounting for the mechanisms at work is important not only for this particular explanation of modal knowledge as an evolutionary byproduct. Generally, when we wish to explain a kind of knowledge from the reliability of the corresponding belief-forming processes, it is not enough to give some reason, evolutionary or otherwise, for why these processes should be reliable. We also need to know, at least in principle, that humans could have cognitive mechanisms that

\textsuperscript{3} See Shope 1983 for an overview of post-Gettier attempts to analyse knowledge.

\textsuperscript{4} See Hawthorne 2004: 7ff. for a recent discussion of lottery cases and reliability, and Goldman 1986 and 1999 for reliabilist approaches in general.

\textsuperscript{5} The \textit{locus classicus} of such an evolutionary explanation is Quine’s account of induction: “creatures inverteately wrong in their inductions have a pathetic but praiseworthy tendency to die before reproducing their kind” (1969: 126). For a more recent discussion of evolutionary explanations in epistemology, see Rescher 1990.

\textsuperscript{6} See Jackson 1982: 133–134 for a different philosophical application of this point.

\textsuperscript{7} On the importance of mechanisms in explanations, see also Railton 1978 and Salmon 1989; the mechanistic paradigm in the philosophy of mind is discussed in Crane 2003.
implement these reliable processes. For instance, it seems that we could come up with a quasi-evolutionary explanation of clairvoyance too. For it would be highly useful to be able to produce true beliefs about things that cannot be perceived through the five senses; for example, one would not have to mount a tedious search for prey if one could tell in advance where it is. However, such an explanation of (alleged) knowledge by clairvoyance would fail, since we have reason to doubt that the human cognitive system could implement processes that tend to yield true beliefs corresponding to the (alleged) knowledge by clairvoyance. To be sure, it may be metaphysically possible that humans should have a reliable faculty for clairvoyance; the point is, rather, that the actual existence of such a faculty seems highly implausible. While knowing the cognitive mechanisms that implement certain belief-forming processes – or at least knowing that humans could have such mechanisms in principle – needs to supplement evolutionary explanations of knowledge and is required for explanations of knowledge as a byproduct of some useful trait, merely identifying the mechanisms does not yield a satisfying explanation of knowledge either. To illustrate this, suppose, for the sake of argument, that our modal beliefs are innate, meaning that people have an innate disposition to acquire certain modal beliefs once they have gained sufficient competence with the concepts involved in those beliefs. It would be interesting to learn this, but it would not be a great advance for epistemology unless we were also given some reason why these innate modal beliefs should tend to be true.

In sum, an explanation of modal knowledge from the reliability of the processes that form our modal beliefs requires that the reliability of those processes be itself explained. This explanation of reliability has two components in turn: first, an account of why the processes that form modal beliefs should be reliable; second, an explanation of the nature of these processes, that is, an account of the cognitive mechanisms that implement them. In other words, we need to answer the questions ‘Why should the processes that form our modal beliefs be reliable?’ and ‘How do these processes work?’ If we succeed in answering both questions, we may achieve an evolutionary explanation of modal knowledge.8

2. Useful Counterfactuals

As we saw in the previous section, an evolutionary explanation of a given phenomenon sometimes requires a detour through other phenomena which are fitness-enhancing and yield the original phenomenon as a byproduct. This section turns to our ability to correctly evaluate counterfactual conditionals as a candidate phenomenon that might yield the reliability of the processes that form our modal beliefs (for short: the reliability of our modal beliefs) as a byproduct.

While the semantics of counterfactual conditionals has enjoyed considerable philosophical interest over the past decades, the epistemology of counterfactuals has been largely neglected.9 It does not seem too difficult to remedy this situation, however, as the epistemology of counterfactual conditionals seems to lend itself to an evolutionary explanation. On the face of it, having a reliable capacity for evaluating counterfactual conditionals is useful. Something goes badly. We reason that if such-and-such had happened, the outcome would have been better. We resolve to bring such-and-such about if we are in a similar situation in the future. When we finally are in a similar situation, we do bring such-and-such about (see Roese and Olson 1995b: 170–172). Succeeding in thus avoiding the bad outcome in the future may have significant practical utility and may even be life-saving. For instance,

8. An anonymous referee has drawn my attention to the fact that Schechter (2010) advocates a model for the explanation of modal and other a priori epistemology that is similar in spirit to the one presented here. He too stresses the importance of explaining reliability and outlines an evolutionary account that answers what he calls the “Operational Question” and the “Etiological Question” about the reliability of the cognitive mechanisms in question (ibid., 444).

9. This has not always been so, at least if we do not conceive of epistemology as restricted to human knowledge. From Suarez to Leibniz, there was an intense debate over the question of how God knows alleged truths like “If the miracles worked in you had taken place in Tyre and Sidon, they would have repented long ago in sackcloth and ashes” (Matthew 11:21). For an overview of this debate, with a focus on Leibniz’s position, see Griffin 1999.
suppose that I am attacked by a tiger while asleep during an expedition and only narrowly escape death. I reason that I would not have been attacked by a tiger if I had lit a campfire, and resolve to light a campfire on similar occasions in the future. Doing so, I avoid tiger attacks for the remainder of the expedition and return home safely.

Besides commonsense evidence, empirical psychology has shown that counterfactuals are useful in learning from mistakes.\(^8\) Experiments have shown that both the resolution to act differently in the future and performance when it comes to action are positively influenced by the prior consideration and acceptance of certain counterfactuals. For instance, the analysis of report forms from NASA’s Aviation Safety Reporting System showed that, after near-collisions, airline pilots who state counterfactuals about what they could have done to avoid the outcome are more likely to resolve to act differently in the future (see Morris and Moore 2000). In repeated flight simulator trials, where they were supposed to land an aircraft under difficult conditions, not only did students who were encouraged to consider such counterfactuals resolve to act differently; their actual performance improved as well (ibid.). Similar results were obtained for less exciting tasks such as anagram solving (see Roese 1994).

Since making correct evaluations of counterfactual conditionals that are involved in learning from mistakes is useful, creatures that tend to evaluate such counterfactuals correctly will have a significant advantage over those who do not; thus ceteris paribus the ability to evaluate these counterfactuals correctly will be selected for. So there is an evolutionary explanation for why the processes that yield our judgements about these counterfactuals should be reliable.

What are the cognitive mechanisms that implement those processes? Unfortunately, little is known about these mechanisms at present. Counterfactual thinking has been studied extensively in psychology (see Mandel, Hilton, and Catellani 2005; Roese and Olson 1995a), but the lion’s share of the studies deal with the consequences of accepting certain counterfactuals, such as emotional effects or conclusions drawn from accepted counterfactuals, and not with the details of the processes that underlie their evaluation.\(^1\) Still, that we have some reliable cognitive mechanisms that underlie our evaluation of counterfactuals does not seem utterly mysterious, unlike, say, the existence of reliable human faculties of clairvoyance. Given that we can provide an evolutionary account of why the processes involved in our evaluation of counterfactual conditionals should be reliable and that the existence of corresponding mechanisms seems plausible, we can explain our reliability in evaluating counterfactual conditionals. At least we can explain our reliability in evaluating those counterfactuals that have practical utility, such as the counterfactual ‘If I had lit a campfire, I would not have been attacked by a tiger’.

How can we extend the explanation of our reliability in evaluating counterfactual conditionals to an explanation of the reliability of our modal beliefs? There is a close logical relationship between the counterfactual conditional and the metaphysical modalities, which is manifest in the following equivalences:

\[
(V) \quad \Box p \text{ if and only if } \neg p \Box \rightarrow p. \text{ That is, } p \text{ is necessary if and only if } p \text{ would be the case if } p \text{ were not the case.}
\]

\[
(V') \quad \Box p \text{ if and only if } \neg p \Box \rightarrow \bot. \text{ That is, } p \text{ is necessary if and only if a contradiction would be the case if } p \text{ were not the case.}
\]

\[
(Q) \quad \Box p \text{ if and only if } \forall q(q \Box \rightarrow p). \text{ That is, } p \text{ is necessary if and only if } p \text{ would be true no matter what was the case.}
\]

The principles (V), (V’), and (Q) are consequences of the Lewisian semantics for counterfactuals.\(^1\) Given the duality of possibility and

\(^{11}\) Byrne 1997, and Johnson-Laird and Byrne 2002, hold that mental models underlie reasoning with counterfactual conditionals. Evans and Over (2004: 123–125) criticise this approach for being too coarse-grained and not accounting for different degrees of “closeness” of possibilities.

\(^{12}\) See Lewis 1973a: 21–24 and Williamson 2005. Lange 2005 and Kment 2006 also characterise modality in terms of counterfactuals; however, they differ from the Lewis-Williamson approach by allowing for false counterpossibles
necessity (□p if and only if¬◊¬p; ◊p if and only if □¬p), we can also derive equivalences of metaphysical possibility with claims involving the counterfactual conditional and thus get, from (V), (V*), and (Q) respectively:

(V*) ◊p if and only if¬(p □→ ¬p). That is, p is possible if and only if it is not the case that p would not be the case if p were the case.

(V**) ◊p if and only if ¬(p □→ ⊥). That is, p is possible if and only if it is not the case that a contradiction would be the case if p were the case.

(Q*) ◊p if and only if ∃q¬(q □→ ¬p). That is, p is possible if and only if there is a proposition such that it is not the case that p would not be the case if this proposition were the case.

Does the intimate logical relationship between counterfactuals and the metaphysical modalities imply an intimate epistemological relationship between these notions? The following two sections examine to what extent these equivalences might afford an explanation of the reliability of our modal beliefs from our reliability in evaluating counterfactuals.

3. Vacuous Truth

The most straightforward way of extending the explanation of our reliability in evaluating counterfactuals to the reliability of our modal beliefs via (V) and (V*) would be to claim that beliefs about necessity, that is, beliefs in instances of the left-hand side of (V), are derived from beliefs in instances of its right-hand side and that beliefs about possibility, that is, beliefs in instances of the left-hand side of (V*), are derived from beliefs in instances of the right-hand side of (V*). Thus, the suggestion is that a belief that p is necessary (B◊p) is derived from a belief that if p were false, p would be true (B¬(p □→ p)), and that a belief that p is possible (B◊p) is derived from a belief that it is not the case that if p were true, p would be false (B¬(p □→ ¬p)).

The problem is that these inferences, and hence (V) and (V*), are not very intuitive. Principle (V) seems to be particularly badly off. People believe (V), if at all, somewhat reluctantly, and mainly out of the theoretical motivation of maintaining an analogy with material conditionals, which are true for false antecedents, and with strict conditionals, which are true for impossible antecedents. This lack of confidence in (V) can be traced to cases where we are more confident about the necessity of a proposition than we are about its counterfactual implication by its own negation. For instance, I am very confident that necessarily Bob is Bob, but only moderately confident that if Bob weren’t Bob, Bob would be Bob. Technically speaking, my degree of belief that necessarily Bob is Bob is high, and significantly higher than my degree of belief that if Bob weren’t Bob, Bob would be Bob. In contrast with such examples, there do not seem to be any cases where our degree of belief in ¬p □→ p is higher than our degree of belief in □p; in the extreme cases, the degrees might be equally low. This asymmetry tells against an epistemological explanation which claims that a belief in □p is inferred from a belief in ¬p □→ p. Degree of belief is never gained in inference; it is preserved or even lost, at least where inferring q from p is the sole basis for believing q. Since often our degree of belief in a proposition of the form ‘p is necessary’ is higher than our degree of belief in the

13. There are alternatives to regarding counterfactuals with impossible antecedents as true. For instance, the conditional □⇒, which is false for impossible antecedents, is briefly explored in Lewis 1973a: 25–26. This result comes at a price, however; for instance, p □⇒ p is not generally true, since it is false if p is impossible. Still, □⇒ may be used to define the metaphysical modalities, since ◊p is equivalent to p □⇒ p.

14. I am putting this claim forward as an empirical observation and emphasise that the qualification that p be the sole basis for believing q is crucial. As it stands, the claim conflicts with standard theories of degrees of belief (see Huber 2011, § 2 for an overview), but it is not uncommon for the human mind to fall short of these theories’ idealisations (see Kahneman, Slovic, and Tversky 1982).
counterfactual implication of \( p \) by its own negation, it cannot generally be the case that a belief in \( \Box p \) is derived from a belief in \( \neg p \rightarrow \Box p \).

Claiming that a belief in \( \Box p \) is derived from a belief in \( \neg p \rightarrow \Box p \) may be a straightforward strategy for explaining the reliability of our modal beliefs from the reliability of our evaluations of counterfactuals via \( (V) \), but its failure does not imply that no alternative strategies are possible. A more sophisticated explanation would claim that beliefs about counterfactuals and modal beliefs have a common source. This would allow that the judgement that \( \neg p \rightarrow \Box p \) is true is made with less confidence than the judgement that \( p \) is necessary, since on this view the latter judgement is no longer derived from the former; if both judgements result from the same cognitive process, the respective degrees of belief may well differ. What might this cognitive process be? The principles \( (V') \) and \( (V'^* ) \) provide a clue. Principle \( (V') \) (and likewise \( (V'^* ) \)) is equivalent to the claim that \( p \) is impossible if and only if \( p \) counterfactually implies a contradiction \( (\neg \Box p \iff p \rightarrow \bot) \). Let us conceive of the evaluation of a counterfactual \( p \rightarrow q \) as the development of the supposition \( p \). Thus, one might conjecture, the process that yields both modal beliefs and evaluations of counterfactuals functions as follows. When evaluating a counterfactual \( p \rightarrow q \), we develop the supposition \( p \). If this developing leads to \( q \), we judge that the counterfactual is true; if it leads to \( \neg q \), we judge that the counterfactual \( p \rightarrow \neg q \) is true. If developing \( p \) leads to a contradiction, we judge that \( p \) is impossible (or equivalently that \( \neg p \) is necessary). If we do not reach a contradiction after we have developed \( p \) for a while, we provisionally judge that \( p \) is possible. If we do not reach a contradiction after we have developed \( p \) for a while and this development leads to \( \neg q \), we provisionally judge that \( p \rightarrow q \) is false. For easy reference, I shall call this the process of joint evaluation.\(^{15}\) Figure 1 is a slightly simplified representation of joint evaluation in the form of a flow chart.

![Figure 1](image-url)  

If we did in fact evaluate counterfactual conditionals and form modal judgements by joint evaluation, the ability to evaluate counterfactuals would bring with it the ability to make judgements of metaphysical modality, and so the reliability of our modal beliefs could be explained as a byproduct of our (evolutionarily useful) reliability in evaluating counterfactual conditionals. Joint evaluation could also avoid the problem that we are often less confident about the counterfactual implication of a proposition by its own negation than we are about the necessity of this proposition. For a proponent of joint evaluation could counterfactuals. See Jenkins 2008 and Peacocke 2011 for further discussion of Williamson’s account.

---

\(^{15}\) Williamson (2007: 141–165) advocates such a model for explaining our ability to make modal judgements as a byproduct of our ability to evaluate
say that in the case where we judge □¬p after having developed p in order to evaluate a counterfactual p □¬ q, we also judge, or are at least disposed to judge, ¬p □¬ ¬p (or p □→ ¬p), although this judgement will be made with less confidence than the judgement of □¬p.\footnote{Similarly, a proponent of joint evaluation might claim that in this case we judge, or are at least disposed to judge, p □¬ ⊥, although this judgement will be made with less confidence than the judgement of □¬p.}

A few remarks on joint evaluation are in order. First, it might seem that joint evaluation cannot explain the production of all our modal beliefs because it presupposes some modal beliefs, namely the belief that contradictions are impossible. This is not the case, however. What is doing the work in joint evaluation is that a proposition that is a contradiction has been reached, not that this proposition is believed to be impossible. That a proposition is a contradiction might be indicated purely by its logical form (such as ‘p and not-p’). If this is granted, joint evaluation can account for our judgement that contradictions are impossible, since, trivially, developing a contradiction leads to a contradiction, viz., itself.\footnote{This suggestion is due to Dorothy Edgington (personal communication).} Second, notice that once we have judged a certain proposition to be impossible in the course of joint evaluation, this proposition can play the role of the contradiction ⊥ in future applications of joint evaluation. For if developing p leads to q, and developing q leads to ⊥, then developing p will also lead to ⊥ by a route that includes the route from p to ⊥.\footnote{This is not to say that developing a supposition is transitive—it had better not be transitive, since counterfactual conditionals are not transitive. We are dealing with a special case, however, since p □→ ⊥ is logically equivalent to p □→ q for any impossible q.} Third, what exactly developing a supposition is stands in need of explanation. Part of developing p may be teasing out the logical consequences of p. But this alone would account for only a limited range of our modal beliefs. It would be unclear, for instance, how we could come to judge that it is impossible for something to be coloured without being spatially extended, since no contradiction can be logically derived from the assumption that something is coloured without being spatially extended. If, on the other hand, developing p and reaching q means judging that if p were the case, q would be the case, joint evaluation would presuppose our ability to evaluate counterfactuals and thus could not explain it. I shall not pursue this worry, however. As we shall see, joint evaluation faces another, more severe, problem.\footnote{Williamson elaborates on the notion of supposing that is at work here at 2007: 143–149.}

Joint evaluation implies that we have a bias towards the possibility of a proposition. When we engage in joint evaluation, we provisionally judge the proposition that is being developed to be possible if the development does not lead to a contradiction after a while. As we keep using joint evaluation, we come to believe an increasing number of propositions to be impossible because developing them leads to a contradiction. This in turn will make it more likely that the development of a new proposition will result in our judging this proposition to be impossible. For the larger our stock of beliefs in impossibilities becomes, the more likely it is that one of these will be encountered in developing a new proposition. (As we just saw, these impossibilities can play the role of ⊥ in joint evaluation.) It may happen that we originally judge p to be possible and later revise this judgement if developing p leads to a newly discovered impossibility, or if p leads to a contradiction after we have developed p more thoroughly than before. By contrast, once we have added a proposition to our stock of impossibilities, joint evaluation does not specify a way for us to revise our belief that p is impossible. So joint evaluation has the consequence that we start out believing few things to be impossible and, correspondingly, many things to be possible; as we build a stock of beliefs in impossibilities, we believe more and more things to be impossible and, correspondingly, fewer and fewer things to be possible.

This prediction, however, conflicts with empirical results. As we grow up, we tend to admit possibilities previously thought to be impossible. For instance, younger children tend to believe it to be impossible that physical laws or social conventions should be
different from how they actually are, while older children tend to believe this to be possible.\(^{20}\) Does this refute joint evaluation? It might be claimed that it does not have to. For it might be possible to explain away the bias towards impossibility in children. One might hold that it is not joint evaluation that is to blame, but rather some sub-process involved in joint evaluation that is somehow deficient in children. For instance, it might be claimed that the children’s process of developing a supposition \(p\) is prone to yield the incorrect judgement that \(p\) leads to a contradiction (and subsequently to yield the judgement that \(p\) is impossible) if \(p\) contradicts certain propositions they know to be actually true. More precisely, it might be claimed, in developing a supposition we hold certain background conditions fixed (see Williamson 2007: 150–153), and children might be prone to hold too many such conditions fixed, which would then easily conflict with the supposition and yield the incorrect judgement that the supposition is impossible. As they grow up, children’s imaginative abilities improve, and they are less and less likely to hold on to too many background conditions. In particular, it might be maintained that the older they are, the more varied their past experiences are; this does not merely lead them to acknowledge the existence of exceptions to false generalisations they have previously held, but also makes them generally less reluctant to suspend belief in what they know to be actually true in the course of developing a supposition.

This reply explains only half of the empirical results. It may explain why children have a bias towards impossibility, but it does not explain why adults lack this bias (or at least have less of it). What is missing is an explanation of how incorrect judgements of impossibility are later revised. As it stands, joint evaluation does not provide such an explanation. It only tells us that initial judgements of a supposition’s possibility may later be overridden, but it does not tell us how judgements of impossibility can ever be revised. And not any suggestion as to how the number of beliefs that such-and-such is impossible is reduced will do. For instance, it would not do to claim that we simply forget them as we grow up. On the face of it, forgetting should be indiscriminate between beliefs in possibilities and impossibilities, thus keeping the ratio between them constant. A simple account of how impossibility judgements are defeated would of course be afforded by the claim that they are revised in the light of later judgements of their possibility. But if this suggestion does not say more about how these possibility judgements are generated, it leaves them unexplained; if it claims that they are the result of joint evaluation, it remains obscure why they should defeat the corresponding impossibility judgements rather than being defeated themselves, as joint evaluation would predict.\(^{21}\)

So more has to be said in order to dispel the apparent conflict between joint evaluation and the impossibility bias in children (and the comparative lack of this bias in adults). This may not refute joint evaluation, which was introduced merely as a rough model of how our judgements of possibility and impossibility relate to our evaluations of counterfactual conditionals and is open to refinement. It does, however, lay the burden of proof on the proponent of joint evaluation.

### 4. Universal Quantification

Instead of taking (V) and (V\(^{\prime}\)) as starting-points for explaining the reliability of our modal beliefs from our reliability in evaluating counterfactuals, we could proceed from (Q). Unlike (V), principle (Q) has the advantage of being highly intuitive. This is evident from

---

\(^{20}\) See Komatsu and Galotti 1986; Miller, Custer, and Nassau 2000. These studies take care to elicit judgements about genuine metaphysical (im)possibility, and not merely judgements about likelihood or truth by phrasing the questions to their subjects in terms of imaginability (Miller, Custer, and Nassau 2000: 388–389) and by using formulations that approximate the possible-worlds criterion for possibility (ibid., 390; Komatsu and Galotti 1986: 414).

\(^{21}\) One might modify joint evaluation so that our belief in the impossibility of a proposition is revised if we empirically come to believe that the proposition is actually true. However, this modification will not explain the many cases where a belief about a proposition’s impossibility is revised without any evidence of the proposition’s actual truth.
the fact that the best strategy for making metaphysical necessity intelligible to non-philosophers normally is to explain that \( p \) is metaphysically necessary just in case it is true \( \text{come what may} \): if pigs were to fly, donkeys were to talk, or anything else were to be the case, \( p \) would still be true. Given the intuitive appeal of (Q), the inference from ‘\( p \) would be true no matter what was the case’ (\( \forall q (q \rightarrow p) \)) to ‘\( p \) is necessary’ (\( \Box p \)) comes naturally to most thinkers. Thus, it looks promising to explain a belief that \( p \) is necessary as inferred from a belief that \( p \) is counterfactually implied by all propositions.\(^{22}\)

There is a worry, however, that being able to reliably evaluate propositions of the form ‘\( p \) is counterfactually implied by all propositions’ is not in the scope of our original explanation of reliability in evaluating counterfactual conditionals.\(^{23}\) On the face of it, what can be explained from our ability to learn from mistakes seems to be our reliability in evaluating ordinary counterfactuals. Having been attacked by a tiger, I judge that I would not have been attacked if I had lit a campfire. This judgement differs in form from, say, the judgement that, whatever had been the case, everything coloured would be spatially extended, where the counterfactual conditional occurs in the scope of a universal quantifier. Thus, the evolutionary explanation may cover the reliability of our beliefs of the form \( B(r \rightarrow s) \), but these differ from beliefs of the form \( B\forall q (q \rightarrow p) \), which are the envisaged sources of our modal beliefs.

However, a closer look at the counterfactual reasoning involved in learning from mistakes reveals that it does in fact cover more than ordinary counterfactuals, such as the counterfactual that I would not have been attacked by a tiger if I had lit a campfire. This counterfactual is about a particular past event: if I had lit a campfire, I would not have been attacked by \textit{that} tiger on \textit{that} occasion. Merely getting this right may enable me to acknowledge my mistake, but not to learn from it. In order to learn from mistakes, I must be able to generalise my insight to similar cases in the future; only that allows me to act differently when endangered by tigers or other predators on future occasions. It seems that when learning from mistakes, we make not just a judgement about a particular counterfactual \( r \rightarrow s \), but also the more general judgement that antecedents similar to \( r \) counterfactually imply consequents similar to \( s \). Formally, we may express this idea by saying that instead of merely judging \( r \rightarrow s \), we judge \( \forall m \forall n (((m \sim r) \& (n \sim s)) \supset (m \rightarrow n)) \), where \( p \sim q \) means that cases where \( p \) is true are sufficiently similar to cases where \( q \) is true.\(^{24}\) If we can explain the reliability of our beliefs of the form \( B\forall q (q \rightarrow p) \), this explanation is not yet as general as an explanation of the reliability of our beliefs of the form \( B\forall q (q \rightarrow p) \) that would be required for explaining the reliability of our modal beliefs.\(^{25}\) But it achieves a significant gain in generality compared to an explanation of the reliability of our beliefs of the form \( B(r \rightarrow s) \).

Still, one might object that the generality gained from generalising the original counterfactual to similar cases does not go far enough. For it might seem that the evolutionarily useful counterfactuals, generalised or not, are still limited to nomologically possible antecedents and consequents, that is, to antecedents and consequents that are metaphysically compossible with the actual laws of nature. Merely being able to correctly evaluate generalised

22. The conditional \( \sim p \rightarrow p \) that is involved in the not-so-intuitive principle (V) is an instance of \( \forall q (q \rightarrow p) \), but by itself this does not affect the claim that the intuitiveness of (Q) suggests that (Q) reflects the cognitive processes that underlie our judgments of necessity (although it does show, again, that we do not live up to the ideals of standard theories of degrees of belief; cf. note 14 above).


24. The judgement \( r \rightarrow s \) can be derived from this more general judgement: since for any proposition, cases where it is true are sufficiently similar to themselves, we can infer \( r \rightarrow s \) by substituting \( r \) for \( m \) and \( s \) for \( n \) in \( \forall m \forall n (((m \sim r) \& (n \sim s)) \supset (m \rightarrow n)) \).

25. On the other hand, an explanation of the former beliefs is more general insofar as it involves counterfactuals with varying antecedents and consequents, while the latter merely involves counterfactuals with varying antecedents.
counterfactuals with nomologically possible antecedents, however, does not suffice for being able to correctly evaluate claims of the form $\forall q (q \implies p)$. For even if we are reliable in our judgements that $p$ is counterfactually implied by all nomologically possible propositions, we may still be vastly wrong in our judgements about whether or not $p$ is counterfactually implied by all nomologically impossible propositions.

This objection should not worry us too much, however, at least if we follow the details of Lewis's (1979, 1981) theory of counterfactuals. Take the counterfactual 'If I had lit a campfire, I would not have been attacked by a tiger'. This counterfactual is non-vacuously true, so according to the Lewisian semantics, some world where I light a campfire and I am not attacked by a tiger is closer to the actual world than all worlds where I light a campfire and I am still attacked by a tiger. Moreover, on Lewis's view, the closest worlds where I light a campfire and I am not attacked by a tiger have the following feature: they match the actual world perfectly until shortly before the actual time of the attack; then a small miracle takes place — perhaps parts of my neural activity suddenly differ from how they actually were — which leads me to light a campfire and thus to avoid an attack.26 This miracle constitutes a breach of the laws of nature as they actually are — a small breach, perhaps, but a breach nonetheless. Thus, despite the fact that both the antecedent and the consequent of the counterfactual 'If I had lit a campfire, I would not have been attacked by a tiger' are nomologically possible,27 the closest worlds where they are true together are nomologically impossible owing to the occurrence of a small miracle.

26. I am assuming here that the counterfactual is assessed under what Lewis (1979: 458) calls the "standard resolution of vagueness", that is, in a standard context.

27. Assuming determinism, neither the antecedent nor the consequent of this counterfactual will be metaphysically compossible with our actual laws of nature plus the prior state of the universe, but this does not contradict their nomological possibility, since there still are possible worlds where either is true together with our laws of nature while the prior state of these worlds differs from that of the actual world.

Countertuals and the Epistemology of Modality

Admittedly, it is not immediately clear how Lewis's theory about when such-and-such counterfactuals are true can be turned into a theory about how we evaluate such-and-such counterfactuals as true. Still, to the extent that his theory is plausible and our evaluations of counterfactuals accord with it, we should not be surprised if the ability to correctly evaluate counterfactuals like 'If I had lit a campfire, I would not have been attacked by a tiger' brought with it the ability to correctly evaluate counterfactuals like 'If I had lit a campfire, a small breach of our laws of nature would have occurred. Thus, the ability to evaluate counterfactuals with nomologically possible antecedents and consequents would bring with it the ability to correctly evaluate counterfactuals with nomologically possible antecedents but nomologically impossible consequents. From the latter, it is a short way to counterfactuals with nomologically impossible antecedents. Generally, $(p \land q) \implies r$ may be inferred from $p \implies q$ and $p \implies r$.28

Assuming that we are able to reason according to this rule, we may infer 'If I had lit a campfire and a small breach of our laws of nature had occurred, I would not have been attacked by a tiger' from 'If I had lit a campfire, a small breach of our laws of nature would have occurred' and 'If I had lit a campfire, I would not have been attacked by a tiger'. Since the conjunction of a nomologically impossible proposition with an arbitrary proposition is still nomologically impossible, the antecedent of 'If I had lit a campfire and a small breach of our laws of nature had occurred, I would not have been attacked by a tiger' is nomologically impossible. Thus, competence

28. Proof. The inference to $(p \land q) \implies r$ from $p \implies q$ and $p \implies r$ is a is logically valid, even though the inference to $(p \land q) \implies r$ from $p \implies r$ alone is not (see Lewis 1973b: 432–433; $p \implies q$ is the 'might' conditional equivalent to $(p \implies q)$ introduced in Lewis 1973a: 21). Since the non-vacuous truth of $p \implies q$ implies $p \implies q$ (ibid.), we get $(p \land q) \implies r$ from non-vacuously true $p \implies q$ and $p \implies r$. Obviously, $p \implies q$ is non-vacuously true if and only if $p \implies r$ is, given that both are true; so should either of $p \implies r$ and $p \implies q$ be vacuously true, the other will be vacuously true as well. In this case, $(p \land q) \implies r$ is vacuously true too; if there are no worlds where $p$ is true, then a fortiori there are no worlds where $p$ and $q$ are true. In sum, the inference from $p \implies q$ and $p \implies r$ to $(p \land q) \implies r$ is logically valid.
with nomologically impossible antecedents and consequents is not too difficult to achieve in principle.

The claim that we are reliable not only about individual counterfactuals but also about generalised counterfactuals of the form \( \forall m \forall n ((m \rightarrow r) \& (n \rightarrow s)) \supset (m \rightarrow n) \) is not confined to the realm of philosophical speculation. For our ability to evaluate generalised counterfactuals can be tested empirically. Suppose that I judge that if \( r \) were the case, \( s \) would be the case, where \( r \) and \( s \) are about certain past events, and generalise this judgement to \( \forall m \forall n ((m \rightarrow r) \& (n \rightarrow s)) \supset (m \rightarrow n) \). Let \( m \) and \( n \) be about certain future events. Suppose further that I judge \( m \sim r \) and \( n \sim s \), and consequently judge \( m \rightarrow n \). It will normally be possible to find out empirically whether or not \( m \) and \( n \) are true. If they both are, the reliability of my judgements about counterfactuals is confirmed, since \( m \rightarrow n \) is true if both \( m \) and \( n \) are. If \( m \) is true while \( n \) is not, \( m \rightarrow n \) is false, and we thus have evidence against one or more of \( r \rightarrow s \), \( \forall m \forall n ((m \rightarrow r) \& (n \rightarrow s)) \supset (m \rightarrow n) \), \( m \sim r \), and \( n \sim s \).

The different premises of our reasoning thus face the tribunal of experience jointly. But it does not follow that we could not in principle narrow down the set of possible culprits to the counterfactual claims, as it might be possible to test the reliability of our judgements about similarity independently.

In sum, the strategy of explaining the reliability of our beliefs about necessity from the reliability of our evaluations of counterfactuals via \( Q \) has a number of advantages: \( Q \) is highly intuitive, and the evolutionary explanation of our ability to evaluate counterfactuals gets us at least very close to claims that a given proposition is counterfactually implied by all propositions, as it can be extended to generalised counterfactuals of the form \( \forall m \forall n ((m \rightarrow r) \& (n \rightarrow s)) \supset (m \rightarrow n) \) and to counterfactuals with nomologically impossible antecedents.

What about the reliability of our beliefs about possibility? \( Q^* \) says that \( p \) is possible if and only if there is a proposition that does not counterfactually imply \( p \)'s negation. It must be conceded that this equivalence is not as intuitive as \( Q \). But reformulating \( Q^* \) may restore intuitiveness. If we introduce the 'might' conditional \( \diamond \rightarrow \) such that \( p \diamond \rightarrow q \) is true if and only if \( \neg \neg(p \square \rightarrow \neg q) \) is (see Lewis 1973a: 21), \( Q^* \) becomes equivalent to the claim that \( \diamond p \) if and only if \( \exists q(q \diamond \rightarrow p) \). Thus, \( p \) is possible if and only if there is some proposition \( q \) such that \( p \) might be true if \( q \) were true. This claim is significantly more appealing than \( Q^* \). Further, 'might' conditionals are often used in arguments about metaphysical possibility. For instance, someone who denied that it was possible for pigs to fly may be challenged by someone else who insists that if pigs had wings and were not as heavy as they actually are, then pigs might fly.

The conditional \( \diamond \rightarrow \) is not merely interdefinable with the 'would' conditional \( \square \rightarrow \); to be reliable in our judgements about 'might' conditionals would also be independently useful. (Henceforth, by a 'counterfactual conditional' \( \text{tout court} \), I shall mean the conditional \( \square \rightarrow \).) For 'might' conditionals often facilitate reasoning with counterfactual conditionals. We saw above that the non-vacuous truth of \( p \square \rightarrow q \) and \( p \rightarrow r \) implies that of \( (p \& q) \square \rightarrow r \). We do not need the counterfactual \( p \square \rightarrow q \) as a premise; the conditional \( p \diamond \rightarrow q \) does the work as well (see note 28). Since \( p \diamond \rightarrow q \) is weaker than the non-vacuous truth of \( p \square \rightarrow q \), we may sometimes know the former without being in a position to know the latter; in this case, it would still be useful to be able to reason to \( (p \& q) \square \rightarrow r \) from \( p \diamond \rightarrow q \) and \( p \rightarrow r \). Now, just as the ability to evaluate single counterfactuals does not directly yield the ability to determine whether or not a given proposition is counterfactually implied by all propositions, the ability to evaluate single 'might' conditionals is not identical to the ability to determine whether or not there is some proposition \( q \) such that a given proposition might be true if \( q \) is. But the step from the former ability to the latter is straightforward, since knowledge of a single conditional of the form \( r \diamond \rightarrow p \) suffices to derive \( \exists q(q \diamond \rightarrow p) \).

29. As was explained in note 24 above, \( r \square \rightarrow s \) can be derived from \( \forall m \forall n ((m \rightarrow r) \& (n \rightarrow s)) \supset (m \rightarrow n) \), so any evidence against the former claim will also be evidence against the latter.
5. Conclusion

To summarise: An evolutionary explanation of our reliability in evaluating counterfactuals looks promising, since the ability to correctly evaluate certain counterfactuals would be practically useful for learning from mistakes and hence be fitness-enhancing. Generally, an evolutionary explanation of a certain feature can be achieved not only if this feature is useful in itself but also if it is the byproduct of some other useful feature. Thus, one strategy for explaining the epistemology of modality is to claim that being reliable in modal judgements is a byproduct of the ability to evaluate counterfactuals. Three equivalences suggest themselves as starting-points for this explanation: the claim that a proposition is necessary if and only if it is counterfactually implied by its own negation, the closely related claim that a proposition is necessary if and only if its negation counterfactually implies a contradiction, and the claim that a proposition is necessary if and only if it is counterfactually implied by all propositions. The first two equivalences lead to the view that judgements of counterfactuals and modal judgements have a common source in a process we called joint evaluation. The third equivalence can be applied to our epistemological problem by claiming that judgements of necessity are derived from universally quantified counterfactuals, and judgements of possibility from existentially quantified ‘might’ conditionals.

Our discussion of how an explanation of the epistemology of counterfactuals might be extended to cover the epistemology of modality has been conditional on assumptions about how judgements about counterfactuals and modal claims are related. Section 3 discussed the assumption that both modal judgements and evaluations of counterfactuals are produced by joint evaluation; the previous section discussed the assumption that modal judgements are derived from universally quantified counterfactuals or existentially quantified ‘might’ conditionals. To some extent, there is philosophical evidence about which of these assumptions is correct. The fact that the equivalence of ‘p is necessary’ with ‘Whatever was the case, p would still be the case’ is highly intuitive, for instance, tells in favour of using this equivalence in an explanation of modal knowledge. However, which assumption is correct cannot be decided by philosophy alone; it also has to take into account results from empirical psychology. For what the cognitive mechanisms underlying certain judgements are is not generally decidable by a priori reflection and introspection alone. Since we do not yet understand the mechanisms very well, what an account of the epistemology of modality will ultimately look like still is very much an open question.

References


31. Thanks to David Etlin, Franz Huber, Tobias Rosefeldt, Timothy Williamson, and two anonymous referees for Philosophers’ Imprint for helpful comments and suggestions.


