A Causal Safety Criterion for Knowledge

Jonathan Vandenburgh

Note: This is a pre-copyedit version of an article published in *Erkenntnis*. Please refer to the final version: https://doi.org/10.1007/s10670-023-00678-3.

Abstract

Safety purports to explain why cases of accidentally true belief are not knowledge, addressing Gettier cases and cases of belief based on statistical evidence. However, numerous problems have been raised for using safety as a condition on knowledge: safety is not necessary for knowledge and cannot always explain the Gettier cases and cases of statistical evidence it is meant to address. In this paper, I argue for a new modal condition designed to capture the non-accidental relationship between facts and evidence required for knowledge: causal safety. I argue that possible errors in belief can be captured by accounting for deviations in causal relationships and that there is a natural way to characterize which causal errors are relevant in an epistemic situation. Using this, I develop a causal analogue to safety, where one's belief in p is causally safe if it is true in all causally relevant worlds where one believes p. Causal safety, I argue, can better explain the cases safety is meant to address and can avoid the arguments raised against the necessity of safety.

Suppose, following an example from Buchak (2014), that your phone was stolen while you were out of the room and that the only two people in the room are Jake and Barbara. You know that men are significantly more likely to steal phones than women are; say, hypothetically, that 9 in 10 stolen phones are stolen by men. On this basis, can you know that Jake stole your phone? It is widely agreed that, even if Jake did steal the phone, you cannot know that he did based just on this statistical evidence.[1] This is because you have insufficient grounds for ruling out the alternative: for all you know, Barbara could have stolen the phone instead of Jake.

A popular condition aiming to capture the robustness to error required for knowledge is the safety condition (Sosa, 1999; Williamson, 2000; Pritchard, 2005). One's belief in p is safe when one's belief could not easily have been wrong, or when p is true in all of the nearest or most similar worlds to the actual world. By taking

[1] Note that Buchak (2014) discusses this case in terms of rational belief rather than knowledge, though the intuition carries over.

safety to be a necessary condition on knowledge, one can purportedly explain why cases where error is salient, such as Buchak's profiling case, are excluded from knowledge. However, safety does not always align with our intuitions about when a belief is sufficiently free from error for knowledge. For example, while statistical evidence about men stealing phones leaves open the possibility that Barbara stole the phone, this possibility may not be nearby or similar to the actual world (Gardiner, 2020). If Jake is a frequent criminal and Barbara is so wealthy that the thought of stealing has never occurred to her, then any possible world where Barbara steals the phone is very distant from the actual world. In this case, your belief is safe, even though intuitively it falls short of knowledge.

This is not an isolated example: many authors have argued that safety cannot explain why cases of accidental (or lucky) true belief are not knowledge (Hiller and Neta, 2007; Pritchard, 2012). Even further, some authors have argued that safety fails as a necessary condition on knowledge (Comesaña, 2005; Kelp, 2009; Bogardus, 2014). In this paper, I argue that these cases arise because of safety's formulation in terms of similarity between possible worlds and develop an alternative theory, causal safety, which can address these critiques of safety.[2] Causal safety builds on a characterization of possible errors in belief as errors in the causal relationships between facts and evidence, where a belief is causally safe if the only possible errors are improbable, non-actual causal errors. Causal safety combines insights from both modal criteria for knowledge like safety and the causal theory of knowledge (Goldman, 1967), and builds on work arguing for causal analyses of counterfactuals over similarity-based theories (Hiddleston, 2005; Pearl, 2009; Briggs, 2012). Causal safety, I argue, can better explain the phenomena safety was introduced to explain, like why knowledge excludes Gettier cases and statistical evidence, and can avoid the counterexamples offered against taking safety as a necessary condition on knowledge.

This paper is organized as follows. In §1, I introduce the safety condition and some of the problems arising for it. In §2, I motivate causal safety as a solution to the problems of safety, and in §3, I introduce causal safety formally using the theory of causal models. In §4, I argue that causal safety can escape the arguments that safety is not necessary for knowledge and cannot adequately address Gettier cases. In §5, I shift focus to cases of statistical evidence, arguing that causal safety offers a robust explanation for why statistical evidence is often insufficient for knowledge, including in the profiling case introduced above.

[2] While I focus on safety in this paper, many of the arguments carry over to related notions utilizing similarity or nearness of possible worlds, such as sensitivity (Nozick, 1981; DeRose, 1995; Enoch et al., 2012) and Lewis's (1996) relevant alternatives theory of knowledge.

1 Knowledge and Safety

Suppose, following an example from Chisholm (1966), that you see what appears to be a sheep in a field. Your visual evidence justifies your belief that there is a sheep, and under normal circumstances, your judgment is correct and you can know that there is a sheep. Gettier cases, however, show that knowledge requires one's evidence to do more than justify a true belief: it must also be appropriately connected with the underlying facts. Imagine, for example, that what you took to be a sheep is actually just a rock, but by chance there is a sheep elsewhere in the field. In this case, your belief that there is a sheep in the field is both true and justified by your evidence, but you do not know that there is a sheep because your evidence is not actually related to the presence of the sheep.

This motivates a further requirement for knowledge beyond justification and truth: if one's belief in p is true merely by chance or by luck, one cannot know p.[3] This non-accidentality condition can explain why Gettier cases are excluded from knowledge, as Gettier cases paradigmatically arise when one's belief is justified and true, but is true only because of some luck or accident disconnected from one's evidence (Gettier, 1963; Zagzebski, 1994). Furthermore, this condition can explain why beliefs based on brute statistical evidence fall short of knowledge, as these beliefs can only be true by chance. For example, in the lottery case (Kyburg, 1961), the odds that a given lottery ticket will lose are very high, but there is no explanation for why the ticket loses beyond luck or chance.

While a non-accidentality condition on knowledge is appealing, it offers little guidance without a more precise account of what it means for a belief to be true by chance or by accident. The most popular characterization of this condition is through safety: a belief is true non-accidentally if it is safe, where belief in p is safe if the agent could not have easily believed p falsely.[4] This is typically explicated in terms of the nearby or most similar possible worlds: belief in p is safe in world w if, in the closest or most similar worlds where one would have believed p, p is true. The safety condition purports to explain why cases of accidentally true belief are not knowledge. In the Gettier case above, there is a nearby world where you believe that there is a sheep based on the rock, but where there is no sheep in the field, so your belief could have easily been false and is therefore unsafe. And in the lottery case, even if your ticket loses in the actual world, there is a nearby world where your ticket won, so believing the ticket will lose based on the low odds of winning is unsafe.

However, further investigation shows safety to be less compelling than it initially appears. Consider the Gettier case discussed above. Here, believing that

[3] See, for example, Unger (1968); Zagzebski (1994); Pritchard (2005).

[4] The non-accidentality of safety is more often formulated as an anti-luck condition; see Pritchard (2005). I will not discuss the subtleties involved in explicating the notions of non-accidentality or epistemic luck; see Schafer (2014); Vogel (2017); Paterson (2020).

there is a sheep based on a rock seems unsafe, as there is a nearby world where you continue believing there is a sheep based on the rock, but where the sheep is gone. This intuition, however, depends on aspects of the case: if the area surrounding the field is inhospitable for sheep, there may be no nearby world where the sheep is outside of the field, rendering your belief safe.[5] This is a common problem for safety: if p could not have easily been false, then one's belief is safe, even if one's evidence for p is not at all connected to the fact that p. Hiller and Neta (2007) and Pritchard (2012) offer further examples in this vein. Imagine someone reads the correct temperature from a broken thermometer, where some further luck guarantees that the broken thermometer is always correct: perhaps a hidden agent sets the broken thermometer to always be correct or the substance measured would explode if the temperature were different. In this case, one's belief is safe, since there is no nearby scenario where your belief is false, but it is not knowledge: a reading from a broken thermometer is insufficiently connected to the actual temperature to ground knowledge.

While these cases challenge the explanatory power of the safety criterion, other cases raise counterexamples to the claim that safety is a necessary condition for knowledge (Comesaña, 2005; Kelp, 2009; Bogardus, 2014). Consider a variant of Comesaña's case: suppose Andy is planning a party at his house and has asked Judy to invite people. Andy does not want John to attend, so he tells Judy to let him know if she informs John of the party location so that he can change it. When John talks to Judy, he introduces himself as Jack, so Andy does not find out that he was invited, meaning that John's belief about the location of the party is true. Here, John knows that the party will be at Andy's house, coming to know it through reliable testimony. However, in a nearby world where John had introduced himself differently, the party would have been moved and John would have had a false belief, making his belief unsafe.

These cases suggest that safety fails to account for the non-accidental connection between one's evidence and the facts necessary for knowledge: there are cases where one's belief is safe, but not appropriately grounded in one's evidence, and cases where one's belief is appropriately grounded in one's evidence, but unsafe. This threatens to undermine both the explanatory power of safety and the claim that safety is necessary for knowledge. The next section motivates a causal safety criterion for knowledge which I argue better captures the non-accidental relationship between one's evidence and the facts needed for knowledge.

2 Motivating Causal Safety

Safety offers an account of the possible errors that block the appropriate connection between facts and belief: belief in p is unsafe if one could have easily made

[5] This criticism of safety is also found in Dutant (2010).

an error about p, or if there is a nearby world where p is false. However, the cases in the previous section show that what happens in nearby possible worlds is not the best way of thinking about the possible errors in belief which preclude knowledge. The errors standing in the way of knowledge can be better understood by analyzing the relationship between facts and evidence more directly. Consider again the case of perceiving a sheep. Here, one's evidence is the perceptual appearance of a sheep, and one infers from this evidence that there is a sheep, as the presence of a sheep is the most likely cause of the evidence. While this inference is justified, it is also fallible. For example, it is possible that there is no sheep at all, and that one's evidence is caused by something other than a sheep, such as a rock, a sheep-like dog, or a hallucination. In causal terminology, these possible deviations are examples of triggering abnormalities, or factors other than a sheep which cause the appearance of a sheep.

Triggering abnormalities can be represented with a causal error term capturing ways in which the world can deviate from what is expected based on causal relationships. Generally, knowledge is compatible with the possibility of causal errors: in normal cases when one perceives a sheep, the possibility of a triggering abnormality causing one's evidence is not sufficient to undermine knowledge. However, there are situations where a causal error term can stand in the way of knowledge. When an error term is activated in the actual world, it can block the appropriate connection between one's evidence and the target proposition, preventing knowledge. For example, in perceiving a sheep, when a triggering abnormality is activated, one's evidence is caused by something other than a sheep, rendering the evidence insufficiently connected to the presence of a sheep to know that there is a sheep. This explains why one cannot know that there is a sheep in the Gettier case: even though the belief is true, the evidence is caused by a triggering abnormality and is therefore disconnected from the fact that there is a sheep. This is part of the non-accidentality of knowledge: even though the belief that there is a sheep is correct, the belief is only correct because a triggering abnormality happened to cause the right kind of misleading evidence. Unlike with the safety criterion, this judgment does not depend on the particular facts of the Gettier case, like whether the sheep would be present in nearby worlds: all inferences that there is a sheep based on a triggering abnormality, like a rock, are disqualified from knowledge.

Requiring that no causal error term intervene between one's evidence and the target proposition, however, is not sufficient to exclude many problematic cases from knowledge. Problems also arise when a causal error blocking the relationship between evidence E and target proposition p does not actually occur, but is likely to occur. In this case, the evidence E is too weak to be definitively linked to p and therefore cannot serve as the basis for knowledge. For example, when perceiving a sheep, if one's evidence is merely a gray-white blob in the distance, the probability that this evidence is caused by something other than a sheep is too substantial to ignore, even if the actual cause of the evidence turns out to be a sheep. Here, the relevant probability is objective rather than subjective: if an error is in fact likely to occur in the evidential situation, then the high probability of error can block knowledge. In addition to excluding cases of weak evidence from knowledge, this condition can explain why one lacks knowledge in cases where errors are extremely likely, such as the barn façade case, and cases where one's evidence is statistically strong, but causally weak, as in the profiling case discussed in §5.

These two conditions motivate the definition of causal safety in the next section: believing p based on E is causally safe when no error terms block the inference of p from E through causal relationships, where an error term stands in the way if either (1) it is likely to be activated or (2) it is activated in the actual world.[6] Following modal accounts of knowledge, this can be interpreted in terms of possible alternatives: belief in p based on E is causally safe in alternative or world w if p is true in all causally relevant alternatives where E is true, where an alternative is causally relevant if all the error terms activated in the alternative are either likely or activated in the actual world.[7] Causal safety captures the intuition that knowledge must be free or safe from error, offering a theory of which errors stand in the way of knowledge.

One consequence of this account of causal safety is that any deviation from the actual world which does not arise from an error term in the causal relationships between p and E is causally relevant. Intuitively, this is part of the non-accidentality of knowledge: if the truth of p depends on the values of some variables one does not have causal evidence for, then even if p is true, it is true because these variables happened to have the correct values by chance rather than by virtue of the evidence E. This excludes from knowledge cases where there is no causal connection between one's evidence and the target proposition, including many cases of statistical evidence. For example, when someone believes that there is a sheep based on a coin toss, whether there is a sheep corresponds to an independent variable in the model one has no causal evidence for, so the conditions for causal safety cannot rule out the alternative that there is no sheep until one

[6] Limiting the domain of worlds to those where you believe p on the basis of E, rather than considering all worlds where you believe p, is analogous to the common 'safe-method' version of safety, where the domain is limited to worlds where you believe p according to the same methods as in the actual world. This is discussed as a condition for sensitivity in Nozick (1981) and Williamson (2000), and Comesaña (2005) and Dutant (2010) discuss the need for such a condition for safety.

[7] One interpretation of this modal notion of causal safety is as a 'normality' approach to knowledge, where knowledge requires that one's belief p is true in all worlds which are normal compared to the actual world. On this interpretation, a world u is abnormal relative to w if it involves activating a low-probability error term which is not activated in w. See the theory of Goodman and Salow (2021), as well as the related normality theory of justification of Smith (2010, 2017).

has strong enough causal evidence.

Thus, causal safety restricts attention to cases where there is a causal connection between E and p. While this may seem restrictive, such a causal connection is present in a wide variety of epistemic circumstances where one has knowledge. Consider the standard sources of knowledge: in perception, memory, and testimony, the fact that p causes one to perceive p, to remember p, or to receive testimony that p. More complex cases of knowledge also involve a causal connection between facts and evidence: knowledge about the future can arise when one's evidence will cause a future event (e.g., buying a plane ticket will cause me to be in Paris) and inference to the best explanation often involves inferring that a theory is the most likely cause of the observed evidence. The causal connection between p and E is closely related to the requirement Goldman (1967) imposes in his causal theory of knowing, where knowledge requires that one's belief in p be causally connected to the world. However, causal safety can handle more cases of knowledge than Goldman's causal theory, as causal safety focuses on causal relationships between variables rather than between one's beliefs and the world, allowing the theory to plausibly extend to more general dependency relations, like those observed in mathematics and ethics.

3 Defining Causal Safety Formally

Causal safety identifies the kinds of errors which are inconsistent with knowledge. These possible errors in reasoning are captured by the error terms in causal relationships. For example, the errors discussed in the previous section for perceiving a sheep are triggering abnormalities, or things other than a sheep which cause the appearance of a sheep. Error terms can be analyzed more formally by representing the causal laws as structural relationships between variables. The theory of causal models used below follows the work of Pearl (2009) in formalizing causality, building on a framework which has been applied extensively in psychology (Glymour, 2001; Sloman, 2005; Gopnik and Schulz, 2007) and the study of language (Hiddleston, 2005; Briggs, 2012).

Variables capture the ways different aspects of the world could be. For example, in perceiving a sheep, there is a variable S representing whether a sheep is present or not and a variable A representing whether there appears to be a sheep. These variables are both binary, i.e., $S = 1$ when there is a sheep and $S = 0$ when there is not a sheep. The value of A depends on the value of S: generally, the appearance of a sheep is caused by a sheep. However, this relationship is not perfectly deterministic. Sometimes, the appearance of a sheep is caused by a triggering abnormality, like a rock or a hallucination. Other times, a sheep does not cause the appearance of a sheep due to an inhibiting abnormality, like when the sheep is hidden or the observer is incapacitated. These abnormalities are captured by error variables: we define U_A to be a binary variable activated when a cause other than a sheep triggers the appearance of a sheep and U'_A to be a binary variable activated when a factor inhibits the appearance of a sheep based on the presence of a sheep. The causal relationship between S and A can then be captured as a structural equation between variables, $A = (S \wedge \neg U'_A) \vee U_A$, which says that there appears to be a sheep when either (1) there is a sheep and nothing inhibits it from appearing that way or (2) something other than a sheep causes the appearance of a sheep. This relationship can also be represented as a causal diagram indicating the influence of S on A:

$$\begin{tikzcd}
S \arrow[d] \\
A
\end{tikzcd}$$

This structure generalizes to other causal relationships through the formalism of causal models. Causal models specify a set of variables and the causal relationships between the variables, represented by structural equations. A causal model $\mathcal{M}$ includes three sets of variables: the exogenous variables U, the error variables $\mathcal{E}$, and the endogenous variables V. It also includes a set of structural equations, $F = \{f_i\}$, such that for each endogenous variable V_i, the structural equation f_i determines the value of V_i given the parents of V_i, PA_i, and the relevant error variables. The parents PA_i are the endogenous and exogenous variables which have causal influence on V_i, and the assignment of parents to the endogenous variables leads to a causal graph $\mathcal{G}$, where an arrow is drawn from V_i to V_j if V_i is a parent of V_j. A causal model can be written as a tuple $\mathcal{M} = (U, \mathcal{E}, V, F)$. In the above case, whether there is a sheep (S) is the exogenous variable, whether there appears to be a sheep (A) is the endogenous variable, U'_A and U_A are the error variables, and the structural equation for A and the causal graph are as described above.

Causal models allow for a convenient representation of epistemic alternatives. Since the endogenous variables are completely determined by the structural equations, all of the facts encoded by a causal model are determined by an assignment to the exogenous and error variables, or a set of values w such that $(U, \mathcal{E}) = w$. The structural equations map values of $(U, \mathcal{E})$ to values of V, so that if $(U, \mathcal{E}) = w$, the structural equations determine the values of the variables in V as $v = F(w)$. The assignments w play the role of epistemic alternatives for causal models, specifying all the different ways things could be consistent with the causal laws; we call the set of all such assignments W. This is analogous to the set of possible worlds from the ordinary conception of safety, and assignments w are sometimes called causal worlds (Pearl, 2009, p. 207). Consider, in the case above, the variable assignment $(S, U_A, U'_A) = (0, 1, 0)$. Since $S = 0$, there is no sheep, and since $U_A = 1$, the structural equation for A requires that $A = 1$, so that there appears to be a sheep based on a triggering abnormality. This represents the epistemic alternative where one sees something that looks like a sheep, but where there is no sheep because the appearance is caused by some other factor, i.e., the case of justified false belief. In this model, there are eight possible worlds corresponding to the eight possible values of S, U_A, and U'_A.

For a causal model to be useful in capturing an epistemic scenario, the propositions of interest must be expressible within the model. Generally, propositions within a causal model are built from assignments to exogenous and endogenous variables: for example, the proposition 'There is a sheep' corresponds to the exogenous variable assignment $S = 1$. Variable assignments correspond to sets of possible worlds: for example, $[U_i = u_i] = \{w \in W : w_i = u_i\}$ and $[V_i = v_i] = \{w \in W : F(w)_i = v_i\}$. Propositions can also include logical combinations of variable assignments: since individual variable assignments are subsets of W, negations, conjunctions, and disjunctions of variable assignments also correspond to sets of possible worlds through set-theoretic complementation, intersection, and union, respectively.

Belief in p based on E is causally safe when one can safely ignore the error terms which stand between E and p in a causal model $\mathcal{M}$. For example, believing that there is a sheep on perceptual evidence is causally safe when one can safely ignore the possibility that one's perceptual evidence is caused by something other than a sheep. As argued in the previous section, causal safety imposes two conditions for one to properly ignore an error variable value: first, the likelihood of the value obtaining must be negligibly small, and second, the value must not obtain in the actual world. This first condition requires some notion of the likelihood that an error term is activated: this can be accomplished by assuming that there is a probability distribution Pr defined over causal alternatives in W, so that Pr assigns a number $\Pr(w) \in [0, 1]$ to each $w \in W$ such that $\sum_{w \in W} \Pr(w) = 1$. Invoking a probability assignment over causal alternatives is common in the causal modeling literature, making the pair $(\mathcal{M}, \Pr)$ a probabilistic causal model (Pearl, 2009, p. 205), and is analogous to defining a probability distribution over possible worlds in semantics and formal epistemology (Hartmann and Sprenger, 2010; Moss, 2015). A probability distribution over causal alternatives determines how likely an error term is to be activated. The probability that an error variable $\mathcal{E}_i$ takes on value ϵ_i is the sum of the likelihoods of each alternative where this requirement is satisfied: $\Pr(\mathcal{E}_i = \epsilon_i) = \sum_{w \in [\mathcal{E}_i = \epsilon_i]} \Pr(w)$. In stipulating that an error term must be improbable given the evidence, I assume that there is a threshold π below which errors are improbable.[8] This allows for a formal definition of

[8] I leave open whether this threshold π is constant across cases or depends on contextual factors. Since causal safety has many features beyond a probability cutoff, this problem is less substantial than it is for other theories. For example, the probability cutoff is not required to satisfy certain properties for causal safety to address problems like the lottery paradox; see Foley (1992); Hawthorne and Makinson (2007); Leitgeb (2014).

causal safety:

Causal Safety: A proposition p is causally safe in causal model $\mathcal{M} = (U, \mathcal{E}, V, F)$ at world $w \in W$ given evidence E and probability distribution Pr if p is true in all alternatives in W' where E is true, where W' is the set of causally relevant alternatives w' from W such that all error assignments $\mathcal{E}_i = \epsilon_i$ in w' satisfy either (1) $\Pr(\mathcal{E}_i = \epsilon_i | E) > \pi$ for some cutoff π or (2) $\mathcal{E}_i = \epsilon_i$ in w.[9]

We can see how this formal account works by reconsidering the case of perceiving a sheep. Here, one believes that there is a sheep, $S = 1$, based on the appearance of a sheep, $A = 1$, and whether one's belief is causally safe depends on whether the error terms inhibit the connection between the evidence and the presence of a sheep. Conditional on the evidence $A = 1$, both error terms are negligibly improbable: the probability that something other than a sheep causes the appearance of a sheep is very low, so $\Pr(U_A = 1|E) < \pi$, and the probability that something prevents the perception of a sheep, conditional on the perception of a sheep, is also negligible, i.e., $\Pr(U'_A = 1|E) < \pi$. Thus, whether one can know that there is a sheep depends only on condition (2), or which error terms are activated in the actual world. In the normal case of perception, where there is a sheep ($S = 1$) which one sees ($U'_A = 0$) and where nothing else causes the perceptual evidence for a sheep ($U_A = 0$), belief that there is a sheep is causally safe. This is because the error assignments $U_A = 1$ and $U'_A = 1$ are both improbable and non-actual, so there are no causally relevant worlds where these error terms are activated: the only causally relevant alternative consistent with one's evidence that $A = 1$ and the structural equation $A = (S \wedge \neg U'_A) \vee U_A$ is the actual world, $(S, U_A, U'_A) = (1, 0, 0)$. On the other hand, in worlds where a triggering abnormality is activated like the case of justified false belief, $(S, U_A, U'_A) = (0, 1, 0)$, or the Gettier case, $(S, U_A, U'_A) = (1, 1, 1)$, belief that there is a sheep is not causally safe. This is because the alternative $(S, U_A, U'_A) = (0, 1, 0)$, where there is no sheep, is causally relevant, since the only activated error term $U_A = 1$ is activated in the actual world, satisfying condition (2) for causal relevance.

It is important to note that the formal definition of causal safety leaves open how the causal model $\mathcal{M}$ and probability distribution Pr are determined. Judgments of causal safety can vary greatly depending on how these are determined, for example, whether these factors arise from subjective beliefs or are objective and agent-independent. While I intend for causal safety to be open to compet-

[9] Note that, because the probability cutoff only applies to a single error term, a proposition which is unlikely to be true can be causally safe. Suppose, following an example from Smith (2017), that 50 people claim they will attend your party, but there is a 10% chance each one does not show up. If each error is negligibly improbable, your belief that all 50 people will attend can be causally safe, even though the odds of this being the case are very low. See also the discussion in Hawthorne and Lasonen-Aarnio (2009) and Williamson (2014).

ing approaches to specifying causal models and probability distributions, I will focus on what I take to be the most plausible account, where the causal model and probability distribution are objective, so that $\mathcal{M}$ represents the actual causal laws and $\Pr(-|E)$ represents an objective chance function conditional on one's evidence. While there are many interpretations of objective chance that can work for causal safety, it may be helpful to think of chance in frequentist terms.[10] For example, in the sheep case, the judgment that $\Pr(U_A = 1|E)$ is low corresponds to the fact that non-sheep causes of sheep appearances make up a small fraction of sheep appearances.

However, even using a frequentist conception of objective chance leaves room for ambiguity in determining the probability distribution. Consider the fake barn case (Goldman, 1976), where an agent correctly judges a building to be a barn based on its appearance, but is only right by chance: the one real barn is surrounded by numerous barn façades, which appear exactly like the real barn, but are not actually barns. Causally, the fake barn case is just another example of perception: we are interested in whether there is a barn (B) based on the appearance of a barn (A), with triggering abnormalities U_A, inhibiting abnormalities U'_A, and structural equation $A = (B \wedge \neg U'_A) \vee U_A$. For the belief that there is a barn to be causally safe, we must be able to ignore the alternatives where $U_A = 1$. Since the agent observes a real barn, no triggering abnormalities are activated in the actual world. Thus, whether this belief is causally safe depends on $\Pr(U_A = 1|E)$, or the probability that a triggering abnormality is activated when there appears to be a barn. The probability of a triggering abnormality is low globally, as it is not often that a barn appearance comes from something other than a barn, but is high locally, as in the local area, the probability that a barn appearance is caused by something other than a barn is very high. In my view, the local interpretation is more compelling, as it is plausible that the probability distribution ought to line up with the frequencies one would observe when sampling from one's environment. On this interpretation, the probability that the observer's evidence is caused by a barn façade in the context is too high to ignore, so the inference that there is a barn from the appearance of a barn is too error-prone to serve as a foundation for knowledge.[11] The fake barn case also highlights how causal safety differs from Goldman's (1967) causal theory of knowing: while the agent's belief is caused by a real barn, this causal connection does not entail that

[10] Other theories of probability that I take to be compelling candidates for use with causal safety are propensity theories of probability and theories of (objective) evidential probability, as in Williamson (2000, Ch. 10).

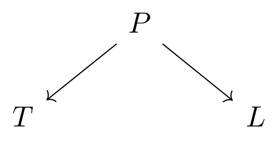
[11] In support of the competing judgment that the probability function takes wide scope in the fake barn case is some experimental evidence that many people attribute knowledge in fake barn cases (Colaco et al., 2014). While I believe that local context can affect probability assignments, as described above, this is not an essential feature of causal safety. In general, how the local environment affects knowledge judgments is a complex question; see Gendler and Hawthorne (2005).

errors in the causal relationship can be properly ignored. Causal safety, unlike Goldman's causal requirement, is a modal property requiring that one's belief be stable across changes to other variables consistent with the causal laws, and this property fails to hold in the fake barn case despite the causal connection between one's evidence and the world.

4 Addressing Arguments against Safety

Like the traditional concept of safety, causal safety is a modal notion: for belief in p to be causally safe based on E in w, p must be true in all the causally relevant alternatives to w where E is true. Causal safety also follows ordinary safety in attempting to capture the non-accidentality of knowledge and the fact that knowledge should be, in some sense, safe from error. However, causal safety makes no reference to the notion of similar or nearby possible worlds: a causal alternative need not be similar to the actual world, and a nearby world need not be causally relevant. For this reason, the arguments against safety discussed in §1 do not apply to causal safety.

Consider first the arguments that safety is not necessary for knowledge, exemplified by the cases of Comesaña (2005), Kelp (2009), and Bogardus (2014). Recall the case inspired by Comesaña (2005): John found out that Andy's party will be at Andy's house from Judy, when he introduced himself as Jack. He almost introduced himself as John, in which case Judy would have told Andy that John was coming and Andy would have moved the party, making John's belief false. Since John's belief could have easily been false, it is not safe, but John seems to know the location of the party based on good evidence appropriately connected to the fact that the party is at Andy's house. Causal safety can explain this. Suppose P represents whether a party is planned for a given location, T represents testimony that the party will be at that location, and L represents whether the party is held at that location. The party being planned at a location ($P = 1$) causes testimony that this is the case ($T = 1$) and causes the party to occur at that location ($L = 1$), though both of these causal relationships can suffer from triggering or inhibiting abnormalities. Thus, the structural equations for the causal relationships are $T = (P \wedge \neg U'_T) \vee U_T$ and $L = (P \wedge \neg U'_L) \vee U_L$ and the causal diagram is as follows:



In Comesaña's example, John believes $L = 1$, that the party will be located at a particular location (Andy's house), based on $T = 1$, (Judy's) testimony that the party will be located there. In the actual world, no error terms are activated, and the party is located at Andy's as planned. For John's belief to be causally safe, $L = 1$ must be true in all causally relevant worlds where $T = 1$. Since none of the error terms are activated in the actual world, the only way an error term can be relevant is if it is sufficiently probable. However, conditional on testimony about the party location ($T = 1$), none of the error terms have high probability of activation: it is rare for people to lie about a party location and rare for people to change the planned party location. This suggests that John's belief is causally safe: none of the error terms are activated or sufficiently probable to be of concern, so the only causally relevant alternative is the actual world where the party is located at Andy's.

Comesaña's example raises problems for safety because the inhibiting error term $U'_L = 1$ is nearly activated: in nearby worlds, Andy intervenes to ensure that the party is not held at his house. However, the fact that U'_L was nearly activated does not mean that the error term is causally relevant. Since the party location is not actually changed, the error term is not activated in the actual world. And since the nearby event is just an outlier possibility, we would not expect it to change the probability $\Pr(U'_L = 1|E)$, which represents the overall likelihood of a party location being determined independent of the plans. Invoking a frequentist approach, we would not expect one party's location nearly being changed to have any effect on the frequency at which parties are changed from their planned locations conditional on testimony.

This strategy for responding to Comesaña's case carries over to other cases of unsafe knowledge. Kelp (2009) and Bogardus (2014), for example, both argue against safety using versions of Russell's (2009) stopped clock case. Suppose, as in Russell's original case, that an agent sees a stopped clock which happens to display the correct time, say 8:22, and comes to form the justified true belief that the time is 8:22. Intuitively, the agent does not know the time, and causal safety can explain this: the actual world is one where a triggering abnormality causes the clock to display the correct time, 8:22, for a reason not connected with the actual time, so it is a causally relevant alternative that the clock displays 8:22, but that the actual time is otherwise. Kelp and Bogardus introduce cases of nearly stopped clocks as counterexamples to safety. Here, the clock the agent uses is perfectly functional and displays the correct time, so the agent knows the time, but there is a nearby possibility that the clock is stopped at 8:22 while the time is not 8:22. Kelp raises the possibility that there is a lazy arch-nemesis who always sets the clock to 8:22, except when it is actually 8:22, and Bogardus supposes that the clock is an atomic clock and there is a nearby alien isotope which could set the clock to 8:22 if it decays. Like in Comesaña's case, the agent's belief is nearly false because an error term could easily have been activated. In both cases, however, no error terms are activated in the actual world, as the clock displays the correct time on the basis of the causal connection between the time and the clock. Furthermore, the fact that the triggering errors were nearly activated is insufficient to make the error terms too likely to ignore: one arch-nemesis tampering with one clock or one almost-compromised atomic clock does not substantially raise the probability of clocks being set by triggering abnormalities rather than the actual time. Thus, as with Comesaña's case above, a nearby possibility of error is not enough to interfere with causal safety, rendering these arguments against the necessity of safety for knowledge inapplicable to causal safety.

While causal safety may escape these counterexamples to the necessity of safety, one might suspect that causal safety also fails as a necessary condition on knowledge, simply for different reasons. Here, I will respond to two lines of objection one might pursue.[12] First, one might think that cases where one can know despite the activation of an error term offer examples of causally unsafe knowledge. Consider the following: you correctly see a sheep in front of you, but behind you someone convincingly imitates a sheep vocalization, thus activating a triggering abnormality for the (auditory) evidence of a sheep. In this case, it seems you can know that there is a sheep based on your good visual evidence, despite the activation of the triggering abnormality. However, not all triggering abnormalities will make one's belief causally unsafe: the only relevant triggering abnormalities are those where activation blocks all causal paths from E to p. In this case, no causal errors block the inference that there is a sheep along the path from one's visual evidence, so your belief that there is a sheep is causally safe. More formally, this case involves sheep S causing visual appearance A and sheep vocalization V according to $A = (S \wedge \neg U'_A) \vee U_A$ and $V = (S \wedge \neg U'_V) \vee U_V$ in world $(S, U'_A, U_A, U'_V, U_V) = (1, 0, 0, 1, 1)$. Even though $U_V = 1$, U_A is neither activated nor likely to be activated, so there are no causally relevant worlds where $U_A = 1$, so the only worlds consistent with evidence $A = 1 \wedge V = 1$ are those where $S = 1$ and $U'_A = 0$, meaning the belief $S = 1$ is true in all causally relevant worlds consistent with the evidence, and thus is causally safe. Thus, a belief does not become causally unsafe whenever a causal error term is activated nearby, but only when this causal error term blocks the inference of p from E in the causal model, a condition much less likely to be consistent with knowledge.

Second, one might think that there are cases of knowledge where the probability of a causal error term being activated is high. Consider a variant of Kelp's case, where a sometimes lazy arch-nemesis always sets the clock to 8:22, except for when it is actually 8:22, when there is about a 1 in 100 chance that he gets lazy and stops interfering. In the case where there is no interference, it seems one can know the time, even though the probability of a triggering abnormality setting the time seems to be .99, and so one's belief appears causally unsafe. However, as in the fake barn case, this verdict depends on the theory of probability used. I have

[12] I would like to thank the anonymous referees for suggesting these cases.

suggested a frequentist interpretation, where probability should line up with the frequency one would observe when sampling from one's (local) environment.[13] On this interpretation, the likelihood of a triggering abnormality altering the clock is still low, since this outcome is unlikely for all the nearby clocks one would sample the probability from.[14] For this probability to be high, one would need to restrict probability to hypothetical trials or propensity of an individual situation, an interpretation which conflicts with the standard goal of using causal models to capture general causal relationships between different settings of variables rather than the actual causal relationships of an individual situation.[15]

While the counterexamples discussed so far offer the most challenging arguments against adopting safety as a necessary condition on knowledge, it is also worth considering a few arguments that safety cannot robustly explain Gettier cases of accidental true belief. For example, in the Gettier case from §1, one's belief that there is a sheep based on a rock can be safe if the sheep elsewhere in the field is also there in nearby possible worlds (such as if the surrounding area is inhospitable), even though one's evidence is not connected to the facts in the right way to support knowledge. This problem does not arise for causal safety: in any version of this Gettier case, the appearance of a sheep is caused by a triggering abnormality, so it is a causally relevant possibility that one's belief is false, regardless of how distant this possibility is. Similar reasoning applies to cases that have been developed to undermine the explanatory adequacy of safety. Imagine, following Hiller and Neta (2007) and Pritchard (2012), that an agent sees a broken thermometer which happens to display the correct temperature, so the agent forms a justified true belief about the temperature which falls short of knowledge. Now suppose that there is no nearby world where the tempera-

[13] For an overview of some criticisms of frequentism, and some alternative approaches to probability, see Hájek (1997, 2019).

[14] This interpretation has an interesting connection to Mortini's (2022) distinction between a 'potentially unfriendly environment' and an 'actually unfriendly environment' in his environment-relative theory of safety.

[15] Note that this point about causal models applying to general cases rather than specific cases may also be relevant to the notoriously difficult case of knowing necessary truths (Dutant, 2010; Roland and Cogburn, 2011). For example, suppose one believes that 47 is prime based on a method M which is in general 50% accurate at identifying prime numbers. Although a causal interpretation is contentious, this could mean that a number being prime (P) has a causal effect on satisfying M, though there is (at least) a 50% chance $M = 1$ is caused by a triggering abnormality rather than the number being prime. Whether the belief that 47 is prime is causally unsafe depends on whether the alternative where 47 is not prime is in the causal model, and if we think of the model as treating P and M as variables that can take on different values for any natural number, then the causal alternative where $P = 0$ should always be in the model, even if the specific number under consideration is (necessarily) prime. This contrasts with the situation for ordinary safety, where it is harder to motivate why there should be a possible world where 47 is not prime, leading authors to modify safety by invoking 'similar propositions' to the proposition that 47 is prime (Williamson, 2009; Hirvelä, 2019), an account which faces further challenges (Zhao, 2022).

ture is otherwise: perhaps, following Hiller and Neta, the thermometer measures a volatile substance which would explode if the temperature were different, or, following Pritchard, there is a hidden agent ensuring that the actual temperature matches the temperature displayed by the broken thermometer. Since there is no nearby world where the actual temperature deviates from the displayed temperature, the agent's belief is safe, even though it falls short of knowledge. The agent's belief, however, is not causally safe: the temperature display is caused by a triggering abnormality rather than the actual temperature, so it is a causally relevant alternative that the actual temperature differs from the displayed temperature, even if such an alternative is distant from the actual world based on other considerations.

Again, one may suspect that causal safety will also fail to address some Gettier cases, particularly in light of Zagzebski's (1994) argument that Gettier cases are inescapable.[16] Such cases may arise if, unbeknownst to the agent, there is a causal structure guaranteeing causal safety in a way that seems inappropriate for knowledge. Consider a variant of Pritchard's thermometer case, where you own a thermometer that you think works normally, but the thermometer reading is actually controlled by a hidden agent who measures the temperature another way and sets the thermometer to the correct temperature. In this case, one might think the hidden agent is part of the true causal structure instead of a triggering abnormality and, provided the agent is so reliable that the odds of a misaligned thermometer are sufficiently low, one's belief can be causally safe merely by chance. However, this verdict depends on contentious modeling decisions: it is not clear that the relevant causal structure should represent the individual thermometer rather than how thermometers operate in general, and even if the agent is included, one might want to include something like the agent's intentions as an independent variable, which would block causal safety. Regardless of how far causal safety can go as a necessary anti-Gettier condition for knowledge, causal safety avoids the cases that problematize safety and, I argue, better accords with our intuitions for when belief is sufficiently grounded in evidence to confer knowledge.

5 Causal Safety and Statistical Evidence

So far, the discussion has focused on Gettier cases, especially variations of Gettier cases designed to pose problems for the safety criterion for knowledge. However, as discussed in §1, non-accidentality conditions on knowledge are also useful for addressing issues which arise for belief based on statistical evidence. In particular, a non-accidentality condition should be able to explain why very strong statistical evidence can be insufficient for knowledge, as in the lottery case where one cannot

[16] Note, however, that Zagzebski excludes conditions that entail truth from her argument, and the causal safety of p guarantees the truth of p since the actual world is always causally relevant.

know that a given lottery ticket will lose simply because the odds of it losing are exceedingly high.

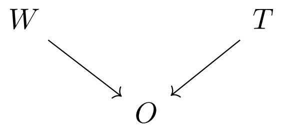
Safety purports to explain why one cannot know based on statistical evidence. In cases where one has strong statistical evidence for p, one's evidence is compatible with some chance that p is false, and these possible worlds where p is false are often no further or more distant from the actual world than worlds where p is true. Consider the lottery case: since tickets are drawn randomly, it could easily be the case that the chosen ticket (or any other ticket) wins, so the belief that a given ticket will lose is not safe (Pritchard, 2005). However, just as problems arise for using safety to explain why Gettier cases are excluded from knowledge, problems arise for applying safety to cases of statistical evidence. Consider the Blue Bus case (Thomson, 1986; Wells, 1992; Enoch et al., 2012): a bus crash causes damage, and one knows that 80% of buses in a city are operated by the Blue Bus Company. Based just on this statistical evidence, one cannot know that the Blue Bus Company is responsible, as the evidence clearly leaves open the possibility that another bus company caused the damage.[17] However, there are situations where one's belief might nevertheless be safe. For example, if no other bus companies were operating near the site of the crash, or if the driver of the bus was uniquely accident-prone of all the bus drivers in the city, then the Blue Bus Company may have caused the crash in all of the nearby possible worlds.

Causal safety offers a more robust explanation for why statistical evidence fails to confer knowledge. The only errors causal safety allows as compatible with knowledge are error terms in causal relationships. When one has merely statistical evidence for p, the alternatives where p is false often correspond to alternative values of independent variables rather than error terms in causal relationships. Since these are not errors arising in the causal connection between p and E, these alternatives are causally relevant, even if they are very improbable.

Consider a causal interpretation of an N-ticket lottery: a player chooses a ticket T from $\{1, ..., N\}$, a winner W is chosen from $\{1, ..., N\}$, and the outcome O is determined from $\{l, w\}$ (lose, win) based, in general, on whether the chosen ticket (T) matches the winning ticket (W). As usual, we can add error terms U_O for situations where one wins even if one's ticket isn't chosen and U'_O for situations where one loses even if one's ticket is chosen. These errors are activated if, say, the lottery is rigged or the game is canceled before a winner is announced. We

[17] While this case is often presented in contexts involving legal liability, I focus only on the epistemic question of whether can know that the Blue Bus Company caused the crash, which may have a different answer from the legal question of whether the Blue Bus Company ought to be held legally liable.

can write the causal graph for the lottery case as follows:



This causal model can explain why someone's belief that their ticket will lose is not causally safe. The relevant evidence E in the situation includes the ticket number one has chosen, $T = t$, as well as statistical information about how the value of the variable W is chosen: for each possible value w of W in $\{1, ..., N\}$, $\Pr(W = w|E) = \frac{1}{N}$. We can assume that both error terms are negligibly improbable, and that the world is such that a different ticket is selected to win ($W = t'$) and neither error term is activated, so the player loses ($O = l$). In evaluating whether the player's belief that $O = l$ is causally safe, we can safely ignore both error terms U_O and U'_O, so the player wins iff $W = t$, which occurs with probability $\frac{1}{N}$. Even though this event is improbable, it remains part of the set of causally relevant alternatives even after removing all of the alternatives with low probability and non-actual error terms. This is because W is an independent variable and no evidence bears directly on which value W takes in the actual world. This means that all possible ticket values, including $W = t$, are causally relevant possibilities. Since $W = t$ is causally relevant, the outcome $O = w$ is causally relevant, preventing one's belief that the ticket will lose from being causally safe.

A similar explanation applies to the Blue Bus Case. Here, since one's evidence is scant, the causal model for the situation only involves a single binary variable B, where $B = 1$ when the Blue Bus Company is responsible and $B = 0$ when another company is responsible. One's evidence includes statistical evidence that the probability the Blue Bus Company is responsible is 80%: $\Pr(B = 1|E) = 0.8$. In this model, there are no error terms one can use to restrict the set of causally relevant worlds, so the set of causally relevant alternatives includes both $B = 1$ and $B = 0$. Thus, the alternative where the Blue Bus Company is not responsible is causally relevant, making one's belief that the Blue Bus Company is responsible causally unsafe, regardless of whether the Blue Bus Company would have been responsible in nearby worlds.

One way of explaining why one cannot have knowledge in these cases is to contrast mere statistical evidence with individualized evidence (Thomson, 1986). On the causal interpretation here, individualized evidence for p is causally connected to p in a causal model, while statistical evidence is not. Suppose, for example, that one reads that ticket t won the lottery in the newspaper and that newspapers misprint lottery winners 1 in N times, or that one learns that the Blue Bus Company caused the crash from eyewitness testimony, which is roughly 80% accurate. In these cases, even though the likelihood one's beliefs are true is the same as in the cases of statistical evidence above, it seems that one's evidence is strong enough for knowledge: one can know that ticket t won by reading it in the newspaper and one can know that Blue Bus Company caused the crash from testimony that this is the case (Enoch et al., 2012). Causal safety explains this: both of these are cases of direct testimonial evidence, where testimony that p is caused by the fact that p, so belief in p is causally safe when triggering abnormalities causing testimony that p are not activated in the actual world and are sufficiently improbable. Thus, belief based on individualized evidence can be causally safe, as the possible errors are error terms in structural causal relationships, while belief based on equally strong statistical evidence is causally unsafe, as the possible errors correspond to different values of independent variables one has no causal evidence for.

However, not all cases of statistical evidence are as clear-cut as the lottery and the Blue Bus cases. Consider the profiling case from Buchak (2014), also discussed in Gardiner (2020): your phone was stolen, and you know that 90% of phones are stolen by men, so you conclude that Jake (rather than Barbara) must have stolen your phone. Intuitively, this case resembles other cases of statistical evidence, as you assume that Jake is responsible while your evidence is insufficient to rule out Barbara. This case is also problematic for the safety account of statistical evidence, as it seems that you cannot know that Jake stole the phone, but there are cases where Jake is responsible in all nearby worlds, such as if Jake is a frequent criminal and Barbara is particularly disposed against stealing. Furthermore, unlike the lottery and Blue Bus cases, the evidence in the profiling case seems more direct: plausibly, being a man has a causal effect on stealing.

However, this causal connection is not strong enough to support causal safety: while being a man is strong statistical evidence that one is responsible for a robbery, it is weak causal evidence.[18] This is because gender is not a key variable explaining theft: while it plays a role, an accurate causal model explaining theft requires additional factors. This is evidenced by research on the causes of delinquency, which focuses on explaining delinquency through personal and social causes, such as lack of self-control (Gottfredson and Hirschi, 1990), personal strain (Agnew, 1992), and factors from social learning, such as attitudes towards the law and imitation of others (Akers, 2011). While gender may not influence delinquency directly, it likely has a causal effect on these personal factors, as would other factors like socioeconomic status, race, and social network. Messerschmidt (2007), for example, argues that masculinity is a factor leading to the kinds of major strain relevant for crime.

This causal structure explains why gender is not strong enough causal evidence for theft: other factors at play are more likely to explain a theft than gender is, so the error term capturing these other factors is very likely to be activated and is

[18] The fact that modal properties like safety can depend on causal structure in cases of statistical evidence is also pointed out in Gardiner (2020).

therefore causally relevant. Consider a simplified causal model: binary variables M, P, and T represent whether someone is a man, whether someone meets the personal/social prerequisites for delinquency, and whether someone commits a theft, respectively, and the diagram for causal influence is as follows:

$$\begin{tikzcd}
M \arrow[d] \\
P \arrow[d] \\
T
\end{tikzcd}$$

As usual, we include both triggering and inhibiting abnormalities, so the structural equations are as follows: $P = (M \wedge \neg U'_P) \vee U_P$ and $T = (P \wedge \neg U'_T) \vee U_T$. In the profiling case, one's evidence E includes the fact that someone committed a theft, $T = 1$, and statistical evidence entailing that $\Pr(M = 1|E) = 0.9$: the probability that the thief is a man is 0.9. For the belief that the thief is a man to be causally safe, $M = 1$ would have to be true in all alternatives once non-actual, low probability errors are removed. Assume that, in the actual world, the thief is a man and none of the error terms are activated: being a man led to the relevant personal and social factors, which led to Jake stealing the phone. Whether $M = 1$ is causally safe depends on how likely the possible error terms are. If we have confidence in theories of delinquency, the error terms $U_T = 1$ and $U'_T = 1$ are negligibly improbable: the factors captured by P offer a causally adequate explanation for a large proportion of thefts, so the likelihood that another cause is at play or that these personal/social causes did not lead to theft, given that a theft occurred, is slim.

However, the error terms for the relationship between personal/social causes and gender are not negligible. This is because, while being a man is highly correlated with the personal and social factors behind delinquency, being a man is not the strongest causal influence on these factors. For many men who exhibit these factors, the real cause is something other than gender, like socioeconomic status or family problems: this corresponds to the case where the triggering abnormality U_P is activated. Thus, even though the probability that someone who commits theft is a man, $\Pr(M = 1|E)$, is high, the probability that some factor other than gender is ultimately causally responsible for this theft, $\Pr(U_P = 1|E)$, is also high. Since the error term U_P is likely to be activated, alternatives where U_P is activated can be causally relevant: in particular, the world where $M = 0$ and the only activated error term is U_P is a causally relevant alternative where E is true, corresponding to the case where someone who is not a man ($M = 0$) commits theft for reasons not related to being a man ($U_P = 1$). Thus, deducing that a man is responsible for a theft based on statistical evidence is not causally safe, as being a man is a causally weak explanation for why someone would commit theft. This explanation also extends to other cases of profiling, as even when properties are strongly correlated with factors like race or gender, the real explanations for the properties often depend on confounding factors which one has no evidence for. Just as in the lottery and the Blue Bus cases, statistical evidence leaves open the values of important factors in a causal model, making it a causally relevant alternative that one's belief is false, even when this is unlikely.

6 Conclusion

Causal safety offers a promising account of the appropriate relationship between evidence and facts required for knowledge. According to causal safety, the only possible errors knowledge is compatible with are those corresponding to causal error terms, and these errors are negligible only if they are both improbable and non-actual. Causal safety can explain why Gettier cases are excluded from knowledge, as Gettier cases arise when one's evidence is disconnected from the proposition of interest, typically through the activation of a causal error term. Causal safety can also explain why knowledge cannot be based on merely statistical evidence, as statistical evidence leaves open key explanatory variables in the causal model and cannot guarantee the truth of the target proposition through the removal of improbable, non-actual errors. Furthermore, causal safety can correctly explain judgments in cases challenging the necessity of ordinary safety. While more work is needed to identify the limitations of causal safety and to better understand the nuances involved in applying formal devices like probabilistic causal models to cases in epistemology, this analysis suggests that the problems identified for ordinary safety do not extend to other modal conditions on knowledge and that tools like causal models may be more useful for the analysis of knowledge than the notion of similarity between possible worlds.

Notes

References

Agnew, R. (1992), 'Foundation for a general strain theory of crime and delinquency', *Criminology* **30**(1), 47–88.

Akers, R. L. (2011), *Social learning and social structure: A general theory of crime and deviance*, Transaction Publishers.

Bogardus, T. (2014), ‘Knowledge under threat’, *Philosophy and Phenomenological Research* **88**(2), 289–313.

Briggs, R. (2012), ‘Interventionist counterfactuals’, *Philosophical studies* **160**(1), 139–166.

Buchak, L. (2014), ‘Belief, credence, and norms’, *Philosophical studies* **169**(2), 285–311.

Chisholm, R. M. (1966), *Theory of knowledge*, Prentice-Hall.

Colaco, D., Buckwalter, W., Stich, S. and Machery, E. (2014), ‘Epistemic intuitions in fake-barn thought experiments’, *Episteme* **11**(2), 199–212.

Comesaña, J. (2005), ‘Unsafe knowledge’, *Synthese* **146**(3), 395–404.

DeRose, K. (1995), ‘Solving the skeptical problem’, *The Philosophical Review* **104**(1), 1–52.

Dutant, J. (2010), Two notions of safety, *in* ‘Swiss Philosophical Preprints’, pp. 1–20.

Enoch, D., Spectre, L. and Fisher, T. (2012), ‘Statistical evidence, sensitivity, and the legal value of knowledge’, *Philosophy & Public Affairs* **40**(3), 197–224.

Foley, R. (1992), ‘The epistemology of belief and the epistemology of degrees of belief’, *American Philosophical Quarterly* **29**(2), 111–124.

Gardiner, G. (2020), ‘Profiling and proof: Are statistics safe?’, *Philosophy* **95**(2), 161–183.

Gendler, T. S. and Hawthorne, J. (2005), ‘The real guide to fake barns: A catalogue of gifts for your epistemic enemies’, *Philosophical Studies* pp. 331–352.

Gettier, E. L. (1963), ‘Is justified true belief knowledge?’, *Analysis* **23**(6), 121–123.

Glymour, C. N. (2001), *The mind’s arrows: Bayes nets and graphical causal models in psychology*, MIT press.

Goldman, A. I. (1967), ‘A causal theory of knowing’, *The Journal of Philosophy* **64**(12), 357–372.

Goldman, A. I. (1976), ‘Discrimination and perceptual knowledge’, *The Journal of Philosophy* **73**(20), 771–791.

Goodman, J. and Salow, B. (2021), ‘Knowledge from probability’, *arXiv preprint arXiv:2106.11501* .

Gopnik, A. and Schulz, L. (2007), *Causal learning: Psychology, philosophy, and computation*, Oxford University Press.

Gottfredson, M. R. and Hirschi, T. (1990), *A general theory of crime*, Stanford University Press.

Hájek, A. (1997), "Mises redux"—redux: Fifteen arguments against finite frequentism, *in* 'Probability, Dynamics and Causality', Springer, pp. 69–87.

Hájek, A. (2019), Interpretations of Probability, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Fall 2019 edn, Metaphysics Research Lab, Stanford University.

Hartmann, S. and Sprenger, J. (2010), 'Bayesian epistemology', *Routledge companion to epistemology* pp. 609–620.

Hawthorne, J. and Lasonen-Aarnio, M. (2009), 'Knowledge and objective chance', *Williamson on knowledge* pp. 305–312.

Hawthorne, J. and Makinson, D. (2007), 'The quantitative/qualitative watershed for rules of uncertain inference', *Studia Logica* **86**(2), 247–297.

Hiddleston, E. (2005), 'A causal theory of counterfactuals', *Noûs* **39**(4), 632–657.

Hiller, A. and Neta, R. (2007), 'Safety and epistemic luck', *Synthese* **158**(3), 303–313.

Hirvelä, J. (2019), 'Global safety: How to deal with necessary truths', *Synthese* **196**(3), 1167–1186.

Kelp, C. (2009), 'Knowledge and safety', *Journal of Philosophical Research* **34**, 21–31.

Kyburg, H. E. (1961), *Probability and the logic of rational belief*, Wesleyan University Press.

Leitgeb, H. (2014), 'The stability theory of belief', *The Philosophical Review* **123**(2), 131–171.

Lewis, D. (1996), 'Elusive knowledge', *Australasian Journal of Philosophy* **74**(4), 549–567.

Messerschmidt, J. W. (2007), 'Masculinities, crime and', *The Blackwell Encyclopedia of Sociology* .

Mortini, D. (2022), 'A new solution to the safety dilemma', *Synthese* **200**(2), 1–17.

Moss, S. (2015), 'On the semantics and pragmatics of epistemic vocabulary', *Semantics and Pragmatics* **8**, 5–1.

Nozick, R. (1981), *Philosophical explanations*, Harvard University Press.

Paterson, N. J. (2020), 'Non-accidental knowing', *The Southern Journal of Philosophy* **58**(2), 302–326.

Pearl, J. (2009), *Causality*, Cambridge university press.

Pritchard, D. (2005), *Epistemic luck*, Clarendon Press.

Pritchard, D. (2012), 'Anti-luck virtue epistemology', *The Journal of Philosophy* **109**(3), 247–279.

Roland, J. and Cogburn, J. (2011), 'Anti-luck epistemologies and necessary truths', *Philosophia* **39**(3), 547–561.

Russell, B. (2009), *Human knowledge: Its scope and limits*, Routledge.

Schafer, K. (2014), 'Knowledge and two forms of non-accidental truth', *Philosophy and Phenomenological Research* **89**(2), 373–393.

Sloman, S. (2005), *Causal models: How people think about the world and its alternatives*, Oxford University Press.

Smith, M. (2010), 'What else justification could be', *Noûs* **44**(1), 10–31.

Smith, M. (2017), *Between probability and certainty: What justifies belief*, Oxford University Press.

Sosa, E. (1999), 'How to defeat opposition to Moore', *Noûs* **33**, 141–153.

Thomson, J. J. (1986), 'Liability and individualized evidence', *Law & Contemp. Probs.* **49**, 199.

Unger, P. (1968), 'An analysis of factual knowledge', *The Journal of Philosophy* pp. 157–170.

Vogel, J. (2017), 'Accident, evidence, and knowledge', *Explaining Knowledge: New Essays on the Gettier Problem* pp. 117–134.

Wells, G. L. (1992), 'Naked statistical evidence of liability: Is subjective probability enough?', *Journal of Personality and Social Psychology* **62**(5), 739–752.

Williamson, T. (2000), *Knowledge and its Limits*, Oxford University Press.

Williamson, T. (2009), 'Probability and danger', *The Amherst Lecture in Philosophy 4* pp. 1–35.

Williamson, T. (2014), 'Very improbable knowing', *Erkenntnis* **79**(5), 971–999.

Zagzebski, L. (1994), 'The inescapability of Gettier problems', *The Philosophical Quarterly* **44**(174), 65–73.

Zhao, B. (2022), 'A dilemma for globalized safety', *Acta Analytica* **37**(2), 249–261.