Synthese (2022) 200:369
https://doi.org/10.1007/s11229-022-03731-w

ORIGINAL RESEARCH

The safe, the sensitive, and the severely tested: a unified account

Georgi Gardiner[1] · Brian Zaharatos[2]

Received: 31 December 2020 / Accepted: 2 May 2022 / Published online: 30 August 2022

Abstract

This essay presents a unified account of safety, sensitivity, and severe testing. S's belief is safe iff, roughly, S could not easily have falsely believed p, and S's belief is sensitive iff were p false S would not believe p. These two conditions are typically viewed as rivals but, we argue, they instead play symbiotic roles. Safety and sensitivity are both valuable epistemic conditions, and the relevant alternatives framework provides the scaffolding for their mutually supportive roles. The relevant alternatives condition holds that a belief is warranted only if the evidence rules out relevant error possibilities. The safety condition helps categorise relevant from irrelevant possibilities. The sensitivity condition captures 'ruling out'. Safety, sensitivity, and the relevant alternatives condition are typically presented as conditions on warranted belief or knowledge. But these properties, once generalised, help characterise other epistemic phenomena, including warranted inference, legal verdicts, scientific claims, reaching conclusions, addressing questions, warranted assertion, and the epistemic force of corroborating evidence. We introduce and explain Mayo's severe testing account of statistical inference. A hypothesis is severely tested to the extent it passes tests that probably would have found errors, were they present. We argue Mayo's account is fruitfully understood using the resulting relevant alternatives framework. Recasting Mayo's condition using the conceptual framework of contemporary epistemology helps forge fruitful connections between two research areas—philosophy of statistics and the analysis of knowledge—not currently in sufficient dialogue. The resulting union benefits both research areas.

Keywords Statistic inference · Severe testing · Error detection in science · Safety · Sensitivity · Relevant alternatives framework · Deborah Mayo

✉ Georgi Gardiner
georgicloud9@gmail.com

[1] University of Tennessee, Knoxville, USA

[2] University of Colorado, Boulder, USA

1 Introduction

Safety and sensitivity are frequently viewed as rival conditions. Debates rage about which one better characterises central epistemic phenomena such as knowledge and justification, and whether safety or sensitivity better responds to skeptical challenges, diagnoses the inadequacy of base rate evidence for outright judgements about individuals, and plays other explanatory roles. This adversative conception, however, is mistaken. They can only be rivals if they compete for the same roles. This essay motivates that safety and sensitivity can instead be fruitfully understood as playing distinct complementary roles in a broader theory of epistemic support. They can work together to characterise central epistemic phenomena and respond to perennial epistemological questions. The relevant alternatives framework provides a unifying structure in which safety and sensitivity play their mutually supportive roles.

Secondly, this essay suggests the resulting framework can help model Deborah Mayo's conception of statistical inference. Mayo's severe testing condition characterises when a statistical inference is supported by the observed data and provides guidance on how practising scientists should collect and use statistical data in building and testing theories. Mayo's rich and fecund research is not widely discussed in mainstream epistemology. It has a lot to offer, but is currently underappreciated.[1] We hope to bridge the apparent chasm between recent developments in epistemology and Mayo's research in frequentist statistical inference. That is, we bring Mayo's research into dialogue with recent mainstream epistemological theory by highlighting their isomorphisms and connections. A closer union would benefit all parties.

In one sense, this essay is ambitious. It aims to unify safety and sensitivity, whilst integrating a theory of statistical inference into mainstream epistemology. But in another sense the aims are modest. We cannot hope to convince doubtful readers in one essay.[2] We only hope to motivate that these ideas are worth pursuing further. Severe testing should be discussed within mainstream contemporary epistemology because it mirrors, and goes beyond, recent developments in modal epistemology. We aim to propel this process.

In section two we explain Mayo's severe testing condition. The basic idea is that a test is severe to the extent that it would detect an error in the hypothesis if an error were present. In section three we explain safety and sensitivity. We argue that putative problems with safety and sensitive indicate the conditions are best seen as playing distinct and collegial roles in a broader theory. Section four sketches this unified account, and recasts Mayo's severe testing condition within this framework. We thereby marry

[1] We argue that Mayo presents a sensitivity condition and develops how that condition operates in practice. Her project thus offers one of the most sophisticated and detailed sensitivity accounts to date. Yet Mayo's work is not cited in any article or book on sensitivity. This discrepancy is noteworthy and demands remedy. (Roush (2005) cites Mayo (1996), but in a different context and not as a sensitivity theory. Cf. Conor Mayo-Wilson (2018).) Similarly, philosophers of statistics do not yet engage with important developments in mainstream non-formal epistemology, such as modal conditions.

[2] Indeed Gardiner (2017, 2020, forthcoming-b) criticises some applications of safety conditions. These objections mostly concern, for example, whether safety distinguishes knowledge from mere true belief. Such objections do not impugn this essay's aims. Other objections, cast doubt on a pure similarity conception of which error possibilities are relevant (Gardiner, 2020).

recent ideas in philosophy of statistics to recent mainstream epistemological theorising. Section five begins to outline some theoretical fruits of this union. We draw further parallels between the views, and highlight insights from one domain that can inform the other. These overlooked parallels are worth highlighting even if ultimately the views are rejected. Indeed, perceiving the parallels can aid detractors, if objections to one view transfer to the other.

Note this paper uses the term ‘sensitivity’ as used in epistemology. (See section three.) This differs from the term ‘sensitivity’ in statistical analysis, where a test’s sensitivity measures the proportion of true positives that are correctly identified.[3]

2 Severity

Broadly, Mayo’s account of severity is motivated by, as she puts it, *finding things out*.[4] For Mayo, finding things out takes hold in the context of statistical science: how does uncertain evidence bear on the kinds of statistical generalisations often found in the empirical sciences? This question is particularly pressing given the replication crisis that has shaken the empirical sciences, especially psychology. There are multiple causes and diagnoses, such as incentivising novel findings, but not spotlighting replication studies or failures to discover correlations.[5] Mayo argues that a major cause was faulty use of statistical tests caused, in large part, by misapprehensions about how statistical inference works. Researchers employed statistical tests without understanding their epistemic contours, which led to inferential errors.[6]

Mayo focuses on the interface between uncertain evidence—or in the parlance of statistics, data—and the epistemology of inference and evidence.[7] She describes how widespread but flawed uses of statistical data lead to impressive looking but spurious results. Data appear to show correlations and so license rejecting the null hypothesis in favour of some positive finding. But these illusive results are the product of bad methods. Such methods, collectively known as ‘p-hacking’ or ‘data dredging’, includes optional stopping rules and post hoc analysis.[8] A theorist might collect more data until statistically significant results are found, for example, or divide the samples into manifold groups to see which groupings yield statistically significant results.[9] These flawed methods ‘practically guarantee’ (Mayo, 2018, p. 5) that a preferred claim, H, will receive support from the data, even if H is false and unwarranted by the evidence. Just about any data, treated with such flawed methods, can seem to support H. Mayo (2018, p. 5) calls this ‘Bad Evidence, No Test’ (BENT).

[3] On the relationship between epistemic sensitivity and type II errors, see Nozick (1981), Pinillos (forthcoming), and Mayo-Wilson (2018).

[4] Mayo (2018, p. 6).

[5] Ioannidis (2005).

[6] Mayo (2018) and Colling and Szűcs (2018).

[7] This project requires translatory grace between research areas. We return to this.

[8] Other colourful terms include data fishing, data snooping, and data butchery; a less colourful collective term is ‘questionable research practices’ or QRPs.

[9] Head et al. (2015).

Mayo offers a simple diagnosis of these flawed statistical inferences: The hypotheses are seen as supported by the observed data, but they have not been subjected to severe tests. She posits a minimum requirement for evidence[10]:

> **Severity requirement (weak)** One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false. If data x agree with a claim C but the method used is practically guaranteed to find such agreement, and had little or no capability of finding flaws with C even if they exist, then we have bad evidence, no test (BENT). (Mayo, 2018, p. 5.)

The data accord with the hypothesis but, Mayo underscores, this does not mean the hypothesis is well supported by the data.[11] If data dredging is used, finding a fit is practically guaranteed. Crucially for the connection to modal epistemology, the severity requirement is understood subjunctively: in BENT cases, were the hypothesis (claim C) false, the data would still fit with claim C. In the parlance of epistemology, the agreement—the fit that is uncovered between data and hypothesis—is insensitive to claim C.[12] We return to sensitivity in section three.

Weak severity is relatively—although, of course, not entirely—uncontroversial. It is a negative condition that diagnoses what is wrong with some flawed inferences. Namely, the putatively supporting evidence would obtain even if C were false. A maths exam cannot test whether a student is good at maths, for example, if a high result is all but guaranteed. (If, say, the student can receive full credit merely by writing their name on the front.) And, we argue below, weak severity maps onto widely endorsed ideas about sensitivity in epistemology. Mayo also endorses a stronger, positive claim[13]:

> **Severity (strong)** We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing result, x, is evidence for C.

[10] In some places Mayo presents Severity as a requirement on something's being good evidence or 'satisfactory evidence' (Mayo & Miller, 2008, p. 309); elsewhere Severity is presented as a requirement on something's being evidence at all. Additionally, Mayo identifies having (good) evidence for H with having a good *test* of H. She writes,

> Following a practice common to testing approaches, I identify 'having good evidence (or just having evidence) for H' and 'having a good test of H.' That is, to ask whether e counts as good evidence for H [...] is to ask whether H has passed a good test with e. (Mayo, 1996, p. 179.)

Mainstream epistemologists who embrace the project laid out here and in Staley and Cobb (2011)—that is, linking error statistics and mainstream epistemology—will likely select only one of these as the relevant explanandum.

[11] In contrast to Mayo's error statistical approach, orthodox Bayesian approaches to statistics are not concerned with error probabilities (Colling & Szűcs, 2018, pp. 7, 12; Gelman et al., 2019, p. 4).

[12] This aspect of Mayo's view might be readily apparent to epistemologists, who are accustomed to subjunctive formulations. But it is overlooked by some statisticians, who then misunderstand and dismiss Mayo's view. See, for example, Bandyopadhyay et al. (2016, pp. 76–77).

[13] Mayo (2018, pp. 14, 179). As above, Mayo frequently switches between whether something is a '[good] test' or '[good] evidence' for a claim. Epistemologists will likely regard these explananda as distinct.

Strong severity aims to characterise the epistemic value of good tests. A good test is good because were H false, the test would have detected it.[14] For observed data e to support a hypothesis H, on Mayo's view, it does not suffice for e to fit H. In addition, e's fitting H must be a good test of H. A test is good if were H false, the data wouldn't fit H. A maths test is a severe test of a student's maths abilities, for example, if a high score is unlikely unless the student was good at maths.

To better understand severe testing, it is helpful to contrast it with rivals. Performance, probabilism, and probativism are competing views of the role that probability ought to play in statistical inference. Performance views posit that the primary role for probability is to characterise long-run properties of statistical methods. In emphasising the need for low type I and type II error rates, Neyman-Pearson hypothesis testing exemplifies a statistical inference method that adopts a performance view.[15] Probabilism holds that the primary role of probability in statistical inference is to quantify the level of support that evidence lends to a hypothesis; 'level of support' is often cashed out in terms of degrees of belief in the hypothesis. Bayesian inference methods assign probabilities to hypotheses based on the posterior distribution from Bayes' theorem, and thus are examples of methods that adopt probabilism. Probativism, by contrast, claims the primary role of probability in statistical inference is to quantify the degree to which a hypothesis has been 'well-probed'.[16] By centring questions about whether the hypothesis has been subjected to a good test, Mayo's measure of severity exemplifies a probativist approach to statistical inference.

Note that although we contrast these three views to better situate Mayo's account within the broader debate, the taxonomy itself is controversial and the terrain is more complex than this tripartite division suggests. In particular, the categories might be better seen as uses of probabilities, rather than overall statistical philosophies. Given this, one might endorse, for example, Bayesianism, but use probabilities in all three ways.[17]

Mayo's severity criterion (SC) for a good test is[18]:

> **Severity criterion (SC)** There is a very high probability that test procedure T would not yield such a passing result, if H were false.

That is, if H were false, probably the data collected by the test would not fit H as well as the actually observed data e do. Mayo restates SC in terms of the improbability of

[14] Here we begin to translate Mayo's ideas into the conceptual framework of contemporary epistemology. She might say severity characterises *when* a statistical inference is good, and when practitioners should make statistical inferences, yet not see her project as explaining the epistemic value of good tests. But her project illuminates epistemic normativity; it can illuminate *why* an inference is good and what the epistemic force of statistical inference is.

[15] Mayo (2018, p. 13) and Neyman and Pearson (1967).

[16] Mayo (2018, p. 162).

[17] Thanks to an anonymous reviewer for helpful feedback.

[18] Mayo (1996, pp. 178–180), Mayo (2018, pp. 92, 149), and Mayo and Spanos (2011, p. 164). In earlier work Mayo does not use the subjunctive mood to characterise severe tests (Mayo, 1996, p. 180). The subjunctive formulation better accords with categorising the view as an arch sensitivity condition, alongside Nozick (1981), Dretske (1970, 1971), Melchior (2019), and others. Compare Guido (2019), who writes 'successfully checking whether p is true requires using a method that is sensitive with respect to p, i.e. a method that would not indicate that p, if p were false'.

the passing result: There is a very low probability that data obtained by the test would have accorded so well with H, were H false. Putting this together yields,[19]

A hypothesis H passes a severe test T with data x_0 if,

(S-1): x_0 accords with H (for a suitable notion of accordance), and
(S-2): with very high probability, test T would have produced a result that accords less well with H than x_0 does, if H were false or incorrect.

Equivalently (S-2) can be stated,

(S-2*): with very low probability, test T would have produced a result that accords as well as or better with H than x_0 does, if H were false or incorrect.

To illustrate, suppose Ronda the wrestler returns from a month abroad and wants to know whether her weight has changed.[20] She previously weighed 112lbs and hopes to compete in her normal weight class of 110–117lbs. Consider claim H: Ronda gained less than five pounds. Ronda worries that H is false—i.e., that she has gained five or more pounds—but based on the evidence that her jeans still fit, Ronda decides that H is true: she gained less than five pounds. Mayo's severity requirement diagnoses that almost nothing has been done to rule out ways that H might be false. There is a good chance that were H false—i.e., that Ronda has gained more than five pounds—the evidence collected through her method would still accord with H. There are many ways Ronda might have gained more than five pounds without outgrowing her jeans, such as muscle gain.

Suppose instead Ronda weighs herself. The scale reads 113lbs. This method is substantially better at discerning if H is false. If H were false, the method would—with very high probability—generate data that do not accord with H. It is worth emphasising that severity is comparative: Ronda could subject claim H to even more severe testing.[21] She could corroborate her result with a second set of scales, for example. This would help eliminate error possibilities in which the first scale was malfunctioning, and it leaves uneliminated only those error possibilities in which both are malfunctioning. It is possible for H to be false—Ronda has gained more than five pounds—and the data accord with H because both scales malfunction, but this error possibility is very unlikely. Claim H is severely tested, and the test is more severe than if she uses just one scale.

The above illustrates Mayo's severe testing with an intuitive, non-statistical example. In what follows, we illustrate severe testing in the context of statistical inference. Readers who would rather focus on the qualitative conception of severity can skip to

[19] Different theories of 'fit' or 'accordance' can be applied to Mayo's framework. The central idea is that e fits H when H renders e probable. H does not fit e if e is improbable under H (Mayo, 2005, p. 99; 124 fn. 3). See also Mayo (1996, pp. 178–182; 2018, p. 92).

[20] This example is based on Mayo (2018, pp. 14–16; 108).

[21] Similarly, the same method and data can test one claim more severely than another. Holding fixed the above method and results, the claim that Ronda weighs less than 140lbs is more severely tested than the claim that she weighs less than 117lbs. The testing method is better able to find flaws with the claim, were it false. Thirdly, holding fixed the hypothesis and the method, different observed data can be a more or less severe test of the hypothesis. If the scale reports her weight as 110lbs, rather than 113lbs, for example, the method would more severely test claim H.

the final paragraph of this section without impeding their understanding of the rest of the paper.

To illustrate the formalised mathematical model of severity, consider Marilynne, the head of the research and development department at Ames' Appliances. Marilynne suspects that a modification to the motor of their best-selling refrigerator will impact the refrigerator's energy consumption, as measured in kilowatts over a 24-h period.[22] She isn't sure whether the modification will have a positive or negative impact on consumption. As such, she might state the following research hypotheses:

R_0: The motor modification will not impact energy consumption
R_1: The motor modification will impact energy consumption

In order to translate the research hypotheses into a formal statistical test, Marilynne must choose a statistical model. She might reasonably assume—perhaps based on knowledge of the measurement process—that the measurements of refrigerator energy consumption are independent, and well-modelled by a normal (that is, Gaussian) probability model. Under these assumptions, Marilynne randomly selects sixty refrigerators from her production line, and randomly assigns a label 'unmodified' or 'modified' to each. As a result, thirty refrigerators undergo a motor modification and thirty remain unmodified.

Under this model, the research hypotheses can be reformulated into statistical hypotheses. Let m_1 be the mean energy consumption in the population of unmodified refrigerators, and m_2 be the mean energy consumption in the population of modified refrigerators.[23] Marilynne's statistical hypotheses are:

$$\mathrm{S}_0 : m_1 = m_2$$
$$\mathrm{S}_1 : m_1 \neq m_2$$

Assuming the variability in kilowatt measurements are the same in the unmodified and modified populations, the test method for these data and these hypotheses is the pooled t-test, which has test statistic:

$$t = \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

where

- $\overline{x}$ is the sample mean of the unmodified group.
- $\overline{y}$ is the sample mean of the modified group.
- $n_x = x_y = 30$ is the number of units in each group.
- s_p is the pooled standard deviation: $s_p = \sqrt{((n_x - 1)s_x^2 + (n_y - 1)s_y^2)/(n_x + n_y - 2)}$.
- $s_x^2 = \frac{1}{n_x - 1} \sum_{\{i=1\}}^{n} (x_i - \overline{x})^2$ is the sample variance for the unmodified group.

[22] This example is modified from Arnholt (2016).

[23] These populations are theoretical. For example, the population of unmodified refrigerators is the set of all instantiations of their best-selling refrigerator model that the manufacturer will make.

- $s_y^2 = \frac{1}{n_y - 1}\sum_{\{i=1\}}^{n}(y_i - \overline{y})^2$ is the sample variance for the modified group.

Marilynne will fix the significance level[24] to $\alpha = 0.05$, and let t_0 denote the value of t for the data collected in this experiment. Marilynne sets the test rule to be:

> T: whenever $t_0 > 2$ or $t_0 < -2$, where t_0 is the test statistic t for our data, infer S1.[25]

At level α, and for the data collected,[26] $t_0 \approx 2.47 > 2$. Thus, Marilynne can infer S1: that the population means of the groups are different, i.e., $m_1 \neq m_2$. That is, if the modelling assumptions are correct, Marilynne can also infer R_1, that, on average, the motor modification has an impact on energy consumption. She can also use the sign of t_0 to infer which group consumes less energy. Since the denominator of t will always be positive, the numerator controls the sign. Since t_0 is positive, it must be that $\overline{x} > \overline{y}$, which implies that the unmodified group used more energy, and that the modified group did better in terms of energy efficiency.

However, it's not clear *how much* better the modification did in terms of energy efficiency. Suppose that, in order for the modification to be financially feasible, the modification must provide at least a 0.5-kilowatt improvement, on average. Let C: the modification made at least at 0.5-kilowatt improvement on average, or, statistically, $m_1 - m_2 > 0.5$. How severely has C been tested? Traditional hypothesis testing does not provide an answer to this question. However, Mayo's severity does. Given our test T and observations—summarized in $\overline{x}$ and $\overline{y}$—the severity of C is approximately 0.03. Severity is measured on a scale from zero—not severely tested—to one—severely tested. Thus, on Mayo's interpretation, C has not been severely tested. Even though her hypothesis test was 'statistically significant', the claim she actually cares about, C, was not severely tested. Marilynne should thus postpone recommending the modification until further testing. Figure 1 shows how severity would change as a function of the kilowatt improvement. Notice that claims about higher gains in efficiency are associated with a lower severity.

As illustrated in the examples above, degree of severity is not a property of the test method simpliciter. It is a function of a test method, a claim C, and an observed outcome, such as the data collected.[27] In the refrigerator case, severity was a function of the pooled t-test, the claim C—the modification made at least at 0.5-kilowatt improvement on average—and the observed data summarized in t_0. Severity of test is thus partly determined by the content of the tested claim and the evidence available. We return to these features of severity below.

[24] The significance level, also known as the 'size' of a test, is the rate of false positive errors, and should be fixed at a tolerable level before the data are collected. Note that, for simplicity, we omit a power analysis for this test, which should be conducted to guide the selection of the sample size.

[25] Two and negative two are the (approximate) critical values for the t-distribution with $n_x + n_y - 2 = 58$ degrees of freedom at level α.

[26] For the data collected, $\bar{x} = 1.76$, $\bar{y} = 1.54$, $s_x^2 = (0.385)^2$, and $s_y^2 = (0.301)^2$.

[27] Mayo and Spanos (2011, p. 164).

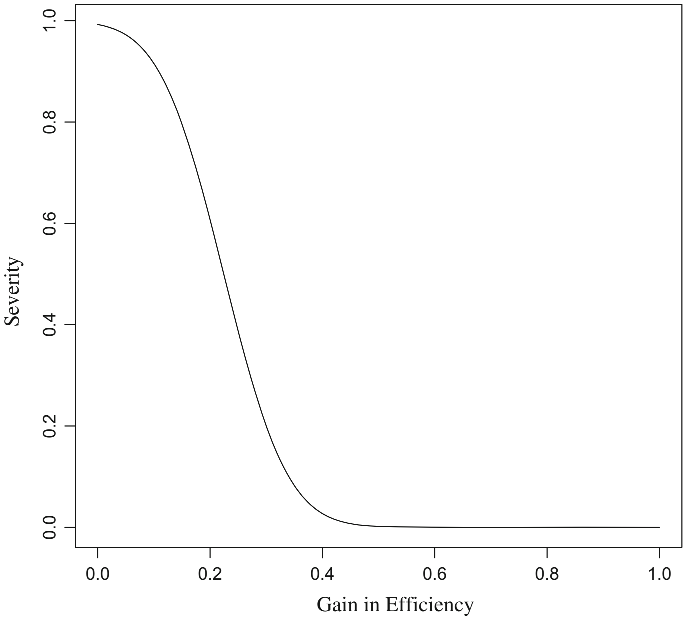

Fig. 1 Severity as a function of the gain in efficiency

3 Safety and sensitivity

Russell looks at a clock, which reads 3 pm. He forms the belief it is 3 pm. And his belief is true. It is 3 pm. Unbeknownst to Russell, however, the clock stopped 24 hours earlier.[28] Intuitively Russell's belief, although true, is not knowledge. A natural explanation for why Russell's belief does not qualify as knowledge appeals to the sheer luckiness of his belief's being correct. He could so easily have been wrong. Had Russell looked at any other time that day he would have formed a false belief. This diagnosis led theorists to posit a safety condition on knowledge.

> **Safety condition on knowledge** S knows p only if S's belief could not easily have been false.

This condition is spelled out in various ways, but the crucial idea is that if S's belief is safe, S would not easily be wrong in a similar case.[29] Duncan Pritchard interprets similarity using a Lewisian possible worlds framework.[30]

> **Pritchard's safety** S's belief is safe if and only if in most nearby possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, and in all very close nearby possible worlds in

[28] Russell (1948). See also Gettier (1963).

[29] Williamson (2000, p. 147). See also Sosa (1999). For a survey, see Rabinowitz (2014). Note that 'could easily have been false' is a controversial idea; its meaning is not straightforward.

[30] Pritchard (2005, 2007, 2009, 2012).

which S continues to form her belief about the target proposition in the same way as in the actual world, the belief is true.

Although Pritchard's formulation divides nearby worlds into two discrete classes—nearby possible worlds and very close nearby possible worlds—this is best understood as a continuum. Closer worlds are more significant for assessments of safety (Pritchard, 2012, p. 255).

The safety condition is marshalled to explain why we cannot know, just by reflecting on the odds, that our ticket did not win a lottery. Although winning is highly improbable, the world need not be very different for the ticket to win, and so a belief formed this way could very easily be false.[31]

The safety condition is externalist. Whether a belief is safe depends on properties of modal space—that is, what in fact would obtain in similar cases—rather than on what the agent believes, or is in a position to know, would obtain in similar cases.

Safety was originally proposed as a condition on knowledge and, accordingly, it is usually presented as a property that an individual person's beliefs can have, based on their total available evidence. But this is not essential to safety's nature.[32] Safety describes a relationship between judgements, their bases, and whether that judgement could easily have been false. Indeed Pritchard (forthcoming) presents safety as an instance of a far more general phenomena: The importance of modal distance from bad outcomes, where false beliefs are just one kind of bad outcome.[33],[34]

Thus safety can be generalised. A judgement is safe iff not easily could the judgement have been wrong, given its basis. The 'basis' can be a body of evidence, epistemic methods, background assumptions, or epistemic character traits. This basis might be socially distributed, formalised, or based on, for example, a restricted subset of evidence, such as legally admissible evidence. The 'judgement' might be a scientific

[31] Whether or not they employ the possible worlds framework, safety accounts rely on some notion of similarity of belief-forming conditions or some comparative notion of whether something 'could easily happen'. Many safety theorists hold that similarity orderings are objective and not interest-dependent. They typically argue, for example, that worlds with smaller physical differences are usually closer—that is, more similar—than worlds with many large-scale physical differences. A world which has different physical laws from this world is more distant than one which exactly resembles our world, except that a few more hemp seeds fell into one porridge bowl this morning. Others deny an overall interest-independent measure of modal nearness or similarity and instead claim, for example, that different interests yield different similarity orderings. Some detractors deny these similarity orderings are clear, cogent, or explanatorily more basic than knowledge, and so reject safety accounts of epistemic phenomena. Theorists who are wholly skeptical of the theoretical foundations of epistemology's modal conditions can instead consider a weaker, albeit still novel, claim advanced in this essay: Safety, sensitivity, and relevant alternatives conditions are usually viewed as rival theories of a target phenomenon, such as knowledge or legal proof. It is fruitful to instead see them as symbiotic conditions that characterise different features of the target phenomena. And Mayo's severity conditions can be fruitfully understood as a sensitivity-based account. Our thanks to an anonymous reviewer for raising this topic.

[32] Our thanks to an anonymous reviewer for pointing out the importance of generalised notions of safety and sensitivity for this project.

[33] Gardiner (2017) argues—against Pritchard—that what is disvaluable is the bad outcome itself, not modal distance from the bad outcome. Safety accounts of the value of knowledge thus face a swamping problem. If the judgement is true, its being safely true contributes no additional value.

[34] Similarly, Gardiner (forthcoming-a, §5) develops a generalised 'relevant alternatives' conception of risk mitigation. The resulting view casts epistemic safeguards, such as corroborating evidence, as structurally isomorphic to prudential safeguards, such as fire alarms. They both have a 'possibility culling' role.

claim, legal verdict, inferential conclusion, or formal institutional finding. Belief is not necessary for some such judgements.[35] This more generalised conception of safety might also help characterise appropriate scientific assertions, question answering, and collectively-held conclusions.[36]

We can similarly adapt Sosa's (1999, p. 142) gloss on safety—'S would believe that p only if it were so that p'—to yield a more generalised formulation. 'The agent would conclude that p only if it were so that p', where the agent might be a group agent, and the 'conclusion' might be a judgement, verdict, assertion, or formal finding.

Some theorists claim an affirmative legal verdict is appropriate only if safe.[37] That is, only if in the nearby worlds—the most similar circumstances—in which the affirmative verdict is reached on a similar basis, that verdict is true. Pritchard claims this condition can explain why bare base rate evidence characteristically does not suffice for affirmative legal verdicts, even when it can render guilt very probable.

The inadequacy of bare base rate evidence for legal verdicts is exemplified by cases like Prisoner.[38]

> **Prisoner** One hundred prisoners exercise in the yard. Security footage reveals that ninety-nine prisoners together attack a guard. One prisoner refuses to participate. Prison officials decide that since for each prisoner it is 99% probable they are guilty, they have adequate evidence to successfully prosecute individual prisoners for assault. They charge Ryan, an arbitrarily selected prisoner in the yard, with assault. A guilty verdict is returned.

Given the evidence, it is highly probable that Ryan rioted. But convicting Ryan on this evidence seems epistemically inappropriate. To explain the epistemic error of convicting Ryan, Pritchard (2015, 2017) argues that legal affirmative verdicts must be safe and the Prisoner verdict is unsafe. He claims that, given the evidence adduced, the verdict against Ryan could easily be false.[39]

[35] For recent surveys of the relationship between belief and scientists' acceptance of their conclusions, see Miller (2014), Elgin (2017), Fleisher (2020), Palmira (2020), and Dang and Bright (2021).

[36] Staley and Cobb (2011, pp. 476–9) also bridge mainstream epistemology and error statistical inference in philosophy of science, including reframing epistemology's internalism–externalism debate to better fit scientific practice. They note that scientific inquiry is typically socially distributed and so is not best characterised by whether beliefs are justified. They instead emphasise the importance of justified *assertions*. Their essay exemplifies how insights from mainstream epistemology and the philosophy of statistical inference can fruitfully inform each other. They forge a groundwork by describing translations and adaptions that can render orthodox mainstream epistemology—with its narrow focus on an individual's belief and knowledge—applicable to science, law, and other socially-extended epistemic enterprises.

[37] Pardo (2018) and Pritchard (2017, forthcoming). For criticism, see Gardiner (2020; forthcoming-b). Gardiner (2019a) surveys modal conditions on legal proof. See especially the section 'Modal Epistemology and the Law'.

[38] Explaining the inadequacy of such evidence for legal proof is known as the 'proof paradox' or 'problem of naked statistical evidence'. For discussion see Cohen (1977), Enoch et al. (2012), Buchak (2014), Blome-Tillmann (2015, 2017), Gardiner (2018, forthcoming-b), Moss (2018b, 2021), and Bolinger (2020). For surveys, see Redmayne (2008), Gardiner (2019a), and Ross (2021).

[39] There might be moral and political reasons against convicting Ryan, but the epistemological project concerns epistemic limits of base rate evidence. These epistemic limits might complement, explain, or be independent from moral and political reasons.

We hold that, *contra* Pritchard, even if legal verdicts are appropriate only if safe, this condition cannot perform all the designated explanatory tasks. When applied to other cases, for example, the safety condition fails to explain the inadequacy of base rate evidence for judgement. Some verdicts qualify as safe simply because p is modally robust. The claim is true in all similar worlds.[40] A person might use poor evidence and reasoning, yet not easily could they be wrong because p is securely true.

Which examples illustrate this is controversial because it depends on similarity orderings. But here is a plausible example: Imagine a rare genetic congenital disease, D. Although rare, if both parents carry D, the offspring will certainly have it. It is genetically determined. In this sense, disease D resembles blood type O, except it is very rare. The modal pattern of disease D appears for any congenital recessive traits that are controlled by a single gene mutation. Cystic fibrosis is a relatively familiar example.[41]

Basil does not know whether his parents have disease D, and he is tested for it. The test is known to have a high true positive rate. That is, the probability that the test shows a positive result, given disease D is present, is high. However, because the base rate of the disease is so low, the probability that Basil has disease D, given a positive result, is low.[42] This fact is explained to Basil. When his test returns a positive result, however, Basil promptly neglects the base rate evidence, and incorrectly calibrates his belief to the high true positive rate; thus he becomes convinced that he carries disease D. Although Basil commits the base rate fallacy, his belief is true. He carries the disease. Given that Basil woudn't exist with different parents and the genetic details of disease D, it is a modally stable feature of his physiology. Basil carries it in all (or almost all) nearby worlds in which he exists.

Safety is ill-equipped to diagnose flaws with Basil's belief. Given the modal stability of his condition, not easily could Basil have falsely believed that he has the disease. If Basil didn't have the disease, he wouldn't be around to form any beliefs at all. Basil's belief is true in all nearby worlds. But his belief is ill-founded; his reasoning was deeply flawed. He committed a statistical fallacy—the base rate fallacy—and his belief is not well-supported by his total available evidence. His only evidence was the positive test result, and many positive test results are inaccurate. Basil should not have been confident that he had the disease, based only on the positive test result.

Basil illustrates that verdicts can be safe simply because the proposition is true in nearby worlds, even if the reasoning used to reach the conclusion is faulty. This

[40] Gardiner (2021b, p. 493) sketches a precursor to the Basil example. Gardiner (2018, p. 184) describes a parallel objection to 'normic' accounts of legal proof. Namely, for 'possible world' analyses of normic support, normal claims are more normically supported just in virtue of being normal. See also Melchior's 'exotitis' and 'Heal the World' examples (Melchior (2019, chapter three)) against safety, Melchior (forthcoming) and Hiller and Neta (2007).

[41] We are grateful to Jelena Aleksic and Ben Martin for their insights.

[42] This follows from Bayes' theorem. Let D denote the event that an individual carries a disease. Let $+$ denote the event that an individual tested positive for the disease. With a true positive rate of $P(+|D) = 0.99$, a false positive rate of $P(+|\neg D) = 0.05$, and a base rate of $P(D) = 0.001$, the posterior probability that Basil has the disease, given the positive test, is $P(D|+) \approx 0.02$.

threatens safety-based explanations of the inadequacy of base rate evidence for verdicts about individuals, including legal verdicts. To see why, consider the following example.[43]

> **Gendered crime** A violent sex crime occurs in a building, and the victim is now deceased. Other than the victim, only Jake (a man) and Barbara (a woman) had access to the building. Jake and Barbara do not know each other well. There is almost no other evidence. The investigator reasons from crime data. She knows that such crimes are almost always committed by men and seldom committed by women. On this basis, she believes Jake is guilty and she charges him with the crime. Her belief is true. Jake did commit the crime.

Jake should not be convicted on this evidence. But in normal versions of this example, a guilty verdict against Jake based on this evidence is modally secure. It is true in all nearby worlds. This is because in nearby worlds where the crime occurred, Jake was the culprit. In these cases, given Jake actually committed the crime, Barbara is not the perpetrator in nearby worlds. Such crimes can be unplanned and opportunistic, but they are not (except in extremely farfetched vignettes) modally like a coin flip or lottery, where the result could easily have been different. The safety condition cannot diagnose why we should not convict Jake on this evidence.

The investigator does not independently know that Jake committed the crime, so she does not know her belief is safe. But this ignorance does not undermine safety because safety is externalist. It depends on how modal space is in fact ordered, and does not directly reflect beliefs about nearby possible worlds.

Given the evidence against Jake, an affirmative verdict is probably true. And, if true, safe. So why is the evidence insufficient?[44] A natural explanation appeals to the importance of detecting error. Were Jake innocent, the available evidence would be identical. The evidence cannot discriminate p from the alternatives.

The capacity to discriminate one possibility from alternatives is of paramount epistemic value. It also has legal value. A state should not convict the defendant unless the evidence adduced can discriminate guilt from innocence. Sensitivity captures this condition.[45]

> **Sensitivity of belief** S's belief that p is sensitive iff if p were false, S would not believe that p

[43] This example is adapted from Buchak (2014)'s iPhone case.

[44] Note that such judgements can have multiple flaws. Finding *a* flaw does not mean identifying *the* flaw. Pritchard contends that safety alone explains the inadequacy of base rate evidence for outright judgment, including legal affirmative verdicts, about individuals. We aver it cannot. Even if safety does some explanatory work, other explanatory conditions are needed. If safety alone cannot explain the relevant epistemic normativity, this raises questions about how safety fits into a broader account. This essay addresses those questions.

[45] Enoch et al. (2012) appeal to sensitivity to explain the inadequacy of base rate evidence for legal convictions. But they deny the epistemic value of sensitivity has legal value. Instead they argue that insensitive verdicts fail to create the proper incentive structures to obey the law. Blome-Tillman (2015) and Gardiner (2018) criticise this suggestion. On sensitivity in epistemology, see Dretske (1970, 1971), Goldman (1976), Nozick (1981), DeRose (1995, 2017), Adams and Clarke (2005), Roush (2007), and Melchior (2019). For a survey, see Melchior (2020).

As with safety, sensitivity is often used to characterise good belief and knowledge, but one can instead give a more generalised conception. A judgement that p is sensitive iff were p false, the agent would not have judged that p. This 'judgement' might be a legal verdict, scientific conclusion, formal finding, news report or similar. The agent might be a group or community. For some such judgements, an individual's believing p is not a central or necessary condition for the judgement that p.

Sensitivity, like safety, is often understood using a Lewisian possible worlds framework: S's true judgement that p is sensitive iff in the nearest possible worlds in which p is not true, S does not judge that p. Applying this more generalised sensitivity condition might explain why one should not convict Ryan with bare base rate evidence. The evidence is wholly insensitive to Ryan's guilt. If he were innocent, the evidence would be the same.

Indeed it is revealing that Pritchard's case for the explanatory power of the safety condition itself illicitly appeals to sensitivity. When confronted with cases like Gendered Crime, Pritchard concedes that an affirmative verdict against Jake is true in all nearby worlds. That is, since Jake committed the crime, he did so in all nearby worlds. But, Pritchard argues, the verdict does not qualify as safe because were the judge to employ this method many times, over a long series of similar cases, she could easily convict an innocent person.[46]

In response: Firstly, in the Gendered Crime vignette the judge only considers Jake's case, and so the worlds where she employs this method many times are modally distant from the original Gendered Crime vignette. It is already a strange case. It would be substantially more strange—even holding fixed that the circumstance happens once—-for many similar cases to occur with the same judge.[47] But safety concerns only nearby worlds. A crucial difference between safety and sensitivity, which allows them to fulfil their respective explanatory roles, is that only nearby possibilities bear on whether a judgement is safe. Distant error possibilities do not undermine a judgement's safety. For sensitivity, by contrast, distant possibilities can make a difference. This is paramount to safety's response to skepticism and undue doubt mongering, its explanation of the possibility of inferential knowledge, and so on. Thus Pritchard's defence should not appeal to distant worlds.

Secondly, even *if* the judge employs the method many times and sequentially convicts each male suspect based on bare base rate evidence, in normal cases that verdict will be true, and thus true in the nearest worlds. And so even if the judge employs the method many times, error only occurs in abnormal cases. If these doubly-distant error possibilities can undermine safety, safety is an extremely demanding condition, and few judgements are safe.[48] Safety is an important epistemic property, but it cannot

[46] See Gardiner (2020, esp. 173–174) for Pritchard's argument.

[47] On the modal surroundings of philosophical vignettes, see Williamson (2007, chapter six). Gardiner (2015a) surveys various subsequent proposals and advocates a normalcy account of modal features of vignettes. The essay claims we should interpret the case's unspecified details to be as normal as possible, given the specified details.

[48] Gardiner (2020) argues that Pritchard's defence of the safety-based explanation inadvertantly leads to skepticism. Some of Pritchard's critics would also maintain that the actual judge's series of other verdicts are irrelevant to whether her target verdict about Jake is safe. This is because the other verdicts are not about

explain the inadequacy of base rate evidence for judgement in cases like Gendered Crime.

The crucial epistemic property missing from the judge's evidence is that *were* Jake innocent, the judge has no way to detect this. She has no safeguard against error. But the safety condition doesn't capture this. The verdict is true in all nearby worlds, so safety is satisfied. The crucial missing property in the Gendered Crime case is sensitivity: The investigator's base rate evidence isn't sensitive.

4 Unification

The parallel between sensitivity and severe testing is apparent.[49] Sensitivity is not a matter of how probable the claim is given the evidence. A judgement can have very high evidential probability, and yet be insensitive. This is exemplified by the lottery, prisoner, and sex crime examples. Instead sensitivity asks 'were the claim false, would this falsity be detectable?' That is, if not p, would the evidence be markedly different? Severe testing likewise focuses on this subjunctive question: If the claim were wrong, would the fit between the favoured hypothesis and the data be notably weaker? And has anything been done so that, were the hypothesis false, the data collected would indicate this falsity? In cases like Prisoner and Lottery, the answer is resoundingly no to both questions.[50]

It is worth emphasising that modal conditions and severe testing were developed to illuminate different things, corresponding to the different guiding aims of theory of knowledge and philosophy of statistical inference. The former characteristically aim to analyse knowledge or justified belief. The latter aims to explain when and how scientists learn from data. Accordingly modal conditions are usually characterised as conditions on belief. Severe testing, by contrast, concerns when evidence suffices to support inferences. Severe testing adherents typically focus on the context of scientific inquiry, including how scientists should audit for errors in their inferences about whether data support a given hypothesis. These differences mean that one cannot directly translate one account into another without modifications. That said, the two domains clearly exhibit—at the very least—illuminating parallels and potential for cross-pollination of research insights. We return to this in section five.

Sensitivity conditions on knowledge face challenges explaining our epistemic position with regard to farfetched and skeptical possibilities. Recall Ronda. She wants to check whether she remains less than 117lbs. Her scale reports 113lbs. Ronda might worry her scales are malfunctioning and so corroborate with separate scales. If she

Footnote 48 continued
Jake and so are wholly different propositions. We can set this objection aside, since the other objections are more forceful and do not turn on precise details of the safety condition.

[49] Although this connection is (by now) apparent to some readers, the subjunctive structure of Mayo's severe testing is overlooked by many researchers (Bandyopadhyay et al., 2016, pp. 76–77; Mayo, 1996, p. 180). Since this paper marries two disparate areas of research, various claims will seem obvious to some readers, but be wholly new to others. We are grateful to Renee Jorgensen for fruitful conversations that crystalised the unified account of safety, sensitivity, severe testing, and the relevant alternatives framework. Together we cannot recall who suggested which ideas, but central parts of the account germinated or coalesced during those conversations.

[50] Enoch et al. (2012). For discussion, see Blome-Tillmann (2015) and Gardiner (2018).

adopts a skeptical attitude, she could remain unconvinced. It is *possible* that both scale readings are wrong, from chicanery or accidental damage. These error possibilities are consistent with her evidence. Ronda can address these possibilities by weighing an object of known weight, such as her dumbbell. If her scales correctly register her dumbbell, this evidence eliminates many error possibilities in which her scales are broken.

Even with this compelling evidence, some error possibilities remain. But they are exceedingly farfetched. It is possible her scales accurately report weights of all other objects, for example, but recently started underreporting Ronda's weight. Some farfetched error possibilities always remain uneliminated. This idea is familiar from the underdetermination of theory by evidence in philosophy of science and some skeptical challenges in epistemology. Such absurd error possibilities can be disregarded in almost any context, of course. They are irrelevant outwith discussions about the contours of skepticism. Mayo articulates a general 'rigged' error possibility for the hypothesis H that Ronda weighs less than 117lbs.[51]

R: Something other than H explains all the data observed so far.

It is consistent with Ronda's observations, no matter how many sets of scales she uses, that H is false and the rigged hypothesis R is true.

The putative problem for sensitivity accounts is that denials of these skeptical error possibilities are insensitive. Consider the non-skeptical claim q: 'it is *not* the case that Ronda's scales accurately report weights of all other objects, but recently started underreporting Ronda's weight'. If q were false then the scale would have some magical but undetectable feature. Her evidence would not be different from Ronda's actual evidence. Were q false, Ronda would continue to believe q. It is characteristic of radical skeptical hypotheses to be consistent with observations. Accordingly their denials, although (presumably) true are insensitive with respect to attainable evidence. They cannot be shown false. No matter what tests Ronda runs, some skeptical possibilities, such as R, remain.[52]

Is this a genuine problem for sensitivity accounts, including Mayo's severe testing? It depends what sensitivity is an account *of*. There is tension between the claims (i.) sensitivity is necessary for knowledge, (ii.) Ronda knows q: 'it is not the case that Ronda's scales accurately report other weights, but recently started underreporting Ronda's weight', and (iii.) Ronda's belief that q is not sensitive. But sensitivity can

[51] Mayo (2018, pp. 15, 108). For error statistical accounts of which error possibilities are relevant, see Staley (2012). See also Mayo and Spanos (2004), Spanos (1999), and Staley (ms; 2008, pp. 400–405). Gardiner (2021b, p. 487; forthcoming-a) describes how error possibilities are divisible; uneliminated error sub-possibilities almost always remain, but uneliminated error possibilities are often sufficiently farfetched to properly ignore. A notable exception is the cogito. The cogito evidence rules out all error possibilities and, accordingly, there is no space for skeptical challenges to take root.

[52] Mayo (2018, p. 108). See Mayo's discussion of prions and Kuru disease (Mayo, 2018, pp. 82–88, 108–110). Strictly speaking, Ronda can run tests for q, which would further chisel the uneliminated remainder. But since this process of addressing increasingly farfetched error possibilities is almost endless, we use the error possibility q to illustrate. Gardiner (2021a, 2021b, §8) describes how the process of ruling out further uneliminated remainders can itself fuel conspiracy theories, undue doubt, mundane skepticism, and other doubt-mongering. This is because uneliminated error possibilities are inevitable and the process of accumulating further corroborating evidence can draw attention to those uneliminated error possibilities. This generates a 'dilemma of engagement' about responding to conspiracy theorists.

play many explanatory roles without being a necessity condition on knowledge.[53] Sensitivity can illuminate, for instance, the nature and value of checking, discriminating, or testing. It can help characterise good tests for whether p, because the results of good tests are sensitive to whether p. Sensitivity can help explain epistemic limits of base rate evidence, and can be required for appropriate assertion, assurance, reactive attitudes, or legal verdicts. Suppose sensitivity characterises checking, for example. The fact that Ronda cannot readily check claim q does not impugn this theory, since we did not antecedently think she could.

An illuminating conception of skeptical challenges is that they attempt to deny us something that we thought we possessed, and that we care about possessing.[54] Perhaps there are some epistemic states, practices, or competences—such as Cartesian certainty about commonplace contingent facts or wholly infallible reasoning, for example—that skeptical reasoning shows we cannot have. If we either do not value those things or should on reflection already realise that we lack them, then conceding the phenomena to skepticism is not perturbing. By contrast there are other states, practices, and competences that we do value, and that we should take ourselves to ordinarily possess. Examples include the legitimacy of our practices of giving and accepting reasons for belief, typically being in a position to assert responsibly, and typically being warranted in trusting our reasoning, perceptions, and memory. Relinquishing these things to skepticism would be a more serious defeat. The skeptical challenge above contends we cannot readily check the denial of farfetched skeptical claims. But it is not a gripping or troubling skeptical challenge unless we antecedently thought we could.

Sensitivity captures key features of epistemic normativity, such as the hallmark of discriminatory abilities and the subjunctive condition of Mayo's severe testing account (that is, S-2 or, equivalently, S-2*). But the role of sensitivity must be situated within a broader account. We must augment sensitivity with a separate and complementary condition that captures which error possibilities can be properly disregarded. For Ronda to test her weight, her evidence must be sensitive. That is, she must be able to discriminate H from various error possibilities. Were H false, her evidence wouldn't fit so well with H. But Ronda need not eliminate every conceivable error possibility, such as the skeptical error possibility R. Skeptical error possibilities like R can be properly ignored, but the sensitivity condition alone cannot capture this feature of epistemic normativity.

A safety condition, by contrast, can help model this feature. If the error possibilities are an 'easy possibility'—if they obtain in nearby possible worlds—then her evidence must address them. If the error possibilities are distant—if the world must be very different for the possibilities to obtain—they can be disregarded. Even though Ronda

[53] This essay does not make claims about the nature of knowledge. Instead it offers a way to unify sensitivity and safety within a relevant alternatives framework and suggests that Mayo's severe testing condition can be fruitfully situated within this framework. Melchior offers a well-developed sensitivity account of checking, rather than knowledge. He argues, for example, failures of closure do not threaten a sensitivity account of checking, even if they threaten a sensitivity condition on knowledge. This is because, Melchior persuasively argues, checking is not closed under known entailment, even if knowledge is. See also Glymour (1980, p. 115); Melchior (2020; forthcoming).

[54] Gardiner (2015b, pp. 42–43; forthcoming-c) develops this conception of skeptical challenges.

cannot rule out farfetched and skeptical error possibilities, Ronda cannot easily be wrong that H. In most contexts, inquirers can conduct ever more tests to rule out increasingly farfetched error possibilities. Appealing to the structure of safety can characterise when inquirers may cease ruling out error possibilities. Inquirers can stop when, given the evidence, not easily could they be wrong.

This is why safety conditions can help explain the lack of knowledge in lottery cases, for example. Lottery case error possibilities—the ticket wins—could easily happen. Relevant alternative theorists provide different overall accounts of which error possibilities are relevant, including sometimes by augmenting a safety-based account with additional conditions, such as whether an error possibility is mentioned or taken seriously by the agent or community. But most, perhaps all, relevant alternative theorists hold that error possibilities that obtain in extremely similar scenarios, or could very easily obtain, are relevant and so must be ruled out.[55]

On the safety-based picture, error possibilities are increasingly distant, where distance corresponds to less 'easy possibilities'. This closeness can be understood in various ways, corresponding to different specifications of the safety condition, such as Pritchard's 'similar possible worlds' view. Safety and sensitivity are not rivals. They play symbiotic roles in a broader account, and the roles can be anchored within a relevant alternatives framework.

Relevant alternatives frameworks were introduced as a condition on knowledge.[56] Lewis (1996) notes that in order to know p, our evidence must eliminate error possibilities. But we need not eliminate every conceivable error possibility. He writes,

> [In order to know p] I may properly ignore some uneliminated [error] possibilities; I may not properly ignore others. Our definition of knowledge requires a *sotto voce* proviso. S knows that p iff S's evidence eliminates every possibility in which not-p—Psst!—except for those possibilities that we are properly ignoring.

More formally,

> **Relevant alternatives condition on knowledge** S knows that p only if S can rule out relevant alternatives to p. Irrelevant error possibilities need not be eliminated.

Characterising irrelevance is contentious. But uncertainty about precisely how to delineate relevance should not precipitate premature dismissal. Like severe testing, the relevant alternatives framework does not currently receive the attention it deserves in

[55] Mayo and Spanos (2004), Spanos (1999), and Staley (2008, pp. 400–405, 2012, ms) might be fruitfully interpreted as error statistical accounts of which error possibilities are relevant, and so brought into dialogue with relevant alternative theories in mainstream epistemology.

[56] Early influential relevant alternatives accounts include Dretske (1970, 1971), Stine (1976), Goldman, (1976), and Lewis (1996). Pritchard (2002) describes how safety conditions resemble relevant alternatives accounts.

epistemology.[57] It can be fruitfully seen as a scaffolding on which different substantive theories can hang. Like safety and sensitivity, the epistemic property need not essentially be about knowledge or belief. The basic framework says that for a claim p and an epistemic standing, such as knowledge or legal proof, some error possibilities must be eliminated and others need not be. One can formulate a more generalised relevant alternatives condition.

> **Relevant alternatives condition, generalised** Claim p is established to an epistemic standard, L, only if the evidence available rules out the L-relevant error possibilities. Irrelevant error possibilities need not be eliminated.

This might be used to model legal standards of proof, such as beyond reasonable doubt, for example.[58]

> **Relevant alternatives condition on 'beyond reasonable doubt'** Claim p is established beyond reasonable doubt only if the evidence adduced rules out the reasonable error possibilities. Irrelevant error possibilities need not be eliminated.

Error possibilities are divisible; they can be rendered into smaller sub-possibilities. An error possibility is addressed by evidence when each sub-possibility is either ruled out by the evidence or is farfetched enough to properly ignore. Theorists endorse rival accounts of what determines remoteness and the disregardability threshold.[59]

This essay does not posit the relevant alternatives condition on knowledge; indeed, it does not require any claims about the nature of knowledge. Instead, we propose that a relevant alternatives framework provides a scaffolding to model Mayo's severe testing and the symbiotic roles of safety and sensitivity. Sensitivity characterises what it means to rule out an error possibility—the evidence is sensitive to the error possibility's obtaining; were the error possibility true, the evidence would reflect this. Safety helps characterise which error possibilities we must eliminate and which we can properly ignore. The resulting framework offers flexibility about precisely how

[57] A relevant alternatives condition frequently appears as a background assumption in theories, yet there is not yet any dedicated overview or survey on relevant alternative theories. There is no *IEP*, *SEP*, *OBO*, or similar. (The only exception is an entry in www.encyclopedia.com.) This fact is telling: This widely-endorsed condition remains under-theorised. For recent accounts employing the relevant alternatives framework, see Rysiew (2006), Gerken (2017), Ichikawa (2017), McKinnon (2013), Ho (2008), Lawlor (2013), Amaya (2015, esp. 525–531), Elgin (2017), Moss (2018a, 2018b, 2021), and Bolinger (2020). Rysiew (2006), Bradley (2014), and Hannon (2015) emphasise the plausibility and universal appeal of a relevant alternatives condition on knowledge and argue the condition is consistent with a wide range of epistemological views. The controversies arise with additional claims—supplementary to the basic relevant alternatives condition—such as contextualism about knowledge attributions.

[58] Gardiner (2019b, forthcoming-a, 2021b) develop 'relevant alternatives' accounts of legal proof, the epistemic force of corroborating evidence, moral encroachment, and when evidence suffices for action, respectively. Gardiner (2021a) applies the relevant alternatives framework to diagnose unreasonable doubt.

[59] See Heller (1989), Lewis (1996), McKinnon (2013), Lawlor (2013), Moss (2018a, 2018b, 2021), Jackson (2018), and Bolinger (2020), for example. For surveys, see Gardiner (2021a, §3; 2021b, §5). An error possibility's remoteness is influenced by probability. Gardiner (2021b; forthcoming-a) argues that—because they can countenance the difference between 'alternative shifting' and 'threshold shifting' mechanisms—relevant alternatives accounts differ fundamentally from standard 'quantifiable probability threshold' accounts of epistemic phenomena.

error possibilities are ordered, reflecting rival accounts of ‘close possibility of error’ and ‘being easily wrong’.

A related proposal is found in Staley (2008, 2012). Staley notes that—once they judge their inferences are warranted—scientists publicly address their scientific claims towards an audience, typically other scientists. According to professional epistemic norms they do so only once they consider themselves ready to defend those claims against challenges. Their peers then present challenges, many of which plumb whether their conclusions are warranted given their evidence. But ‘such challenges are not posed arbitrarily’ (Staley, 2012, p. 30). Only some kinds of challenges are deemed appropriate, namely the ones ‘judged significant’ (ibid.).

Staley characterises systemising which error possibilities are relevant—that is, which challenges are epistemically appropriate—as the ‘most pressing problem’ for the resulting account of the epistemology of statistical inference.[60] He proposes a way to sort relevant from irrelevant error possibilities for severe testers. Severe testing requires the specification of a statistical model. A statistical model articulates the assumptions about the particular statistical characteristics of the data generating process, and so defines the statistical test. This includes, for example, whether the data are independent and identically distributed (IID). On Staley’s view, the relevant error possibilities are those compatible with model assumptions that define the statistical test.[61] He then posits that justifying claims about which hypotheses are supported by the data proceeds by securing the claim against ‘scenarios under which it would be incorrect’. That is: against error possibilities.

One increasingly ‘secures’ the evidence ‘by showing that, given […] one’s epistemic situation, the ways in which one might go wrong can be ruled out, or else make no difference to the evidential conclusion one is drawing’ (Staley, 2012, p. 30). Security is understood as ‘truth across epistemically possible scenarios’ (2012, p. 23). Full security is usually an unreachable ideal. He discusses ways one might increase the security of an inference by either weakening the conclusion or strengthening the evidence base.[62]

One can thus ‘compare and contrast’ Staley’s proposal for which error possibilities are relevant with the many existing ones within epistemology’s ‘relevant alternatives’ literature.[63]

[60] Staley (ms, p. 28).

[61] The adequacy of these model assumptions must itself be tested (2008, pp. 401, 408, 2012). See also Mayo and Spanos (2004) and Spanos (1999).

[62] See Staley (2012, p. 32). The two kinds of strategy are illustrated in Staley (2012, §§4 and 5). Staley (2012, pp. 38–41) illustrates the latter strategy with Mayo and Spanos’s misspecification testing and model re-specification (Spanos, 1999; Mayo & Spanos, 2004). On some relevant alternative accounts, one can also render a judgement justified by changing the judgement’s context so that fewer error possibilities are relevant. On such views, this can happen if, for example, stakes are lowered or one’s community stops taking the error possibilities seriously. These mechanisms are described, but not endorsed, in Gardiner (2021a, 2021b) respectively.

[63] Staley’s (2008, pp. 404–405) account shares suggestive content with Pritchard’s (2002) safety-based approach to understanding relevant alternatives, such as that determining the relevant error possibilities is based on judgements about the ‘most similar’ cases. Its emphasis on the importance of assertion and defence from social challenge finds kin in Austin’s (1946) and Lawlor’s (2013) relevant alternatives accounts.

Controversies about the precise analysis of ‘easy possibility of error’ and demarcating relevant from irrelevant error possibilities does not stymie the severe testing view proposed here. This is because these keystone notions are ineliminable in ordinary thought and talk. Accordingly, few theorists claim they are incomprehensible. Theorists should employ and study such crucial everyday ideas. If mere contentiousness disqualified theorists from using a theoretical posit, furthermore, most research would stall. Indeed the posits of rival accounts of statistical inference, such as priors, are themselves contentious. Lastly, difficult and controversial cases are unlikely to affect the resulting severe testing account because severe testing is rooted in scientific practice, rather than obscure philosophical examples. We thus hope to sidestep questions about how error possibilities are ordered, and we instead emphasise the potential for mutual illumination between relevant alternatives accounts and Mayo’s research about which scientific error possibilities should be eliminated.

Thus we can harness recent epistemological theory to model Mayo’s severe testing. This brings error statistics into fruitful dialogue with developments in mainstream contemporary epistemology. This union is fecund. Mayo’s error statistical view is one of the most advanced and sophisticated sensitivity accounts, and yet isn’t discussed—or even mentioned—by any sensitivity research in epistemology. The barrier is a palisade, not a ha ha: neglect of consilience is mutual.[64] Mayo has independently developed a sensitivity condition without drawing on the resources of contemporary epistemological theory. She has developed a sensitivity account, without perceiving herself as such. Similarly, Staley develops a ‘relevant alternatives’ account without connecting it to existing ‘relevant alternatives’ research.

Mayo’s severe testing provides a highly developed sensitivity account of when and how statistical inferences in scientific practice are sensitive to error possibilities. She provides a panoply of statistical methods for detecting errors. This research can be harnessed by epistemologists. Conversely, recent developments in epistemology can enhance Mayo’s view. Inquirers need not eliminate all error possibilities. Indeed, one couldn’t. Some are disregardably farfetched or skeptical. Scientists must eliminate the ‘easy possibilities of error’. On a safety account, those alternatives that are close possibilities and obtain in nearby possible worlds. The resulting suggestion uses the relevant alternatives framework to unify safety and sensitivity, and situates Mayo’s view within this picture.

5 The fruits of consilience

We close by motivating the project of further unifying these two areas. We highlight some germinal connections between Mayo’s probativist account of statistical inference and recent epistemological theorising. These ideas are embryonic. Rather than provide watertight arguments for claims, we suggest potential avenues for future inquiry. This aims to be simply an invitation to further dialogue; hors d’oeuvres to entice discussants to the table.

[64] A ha ha is a sunken fence that obscures the view in one direction, but not the other. A palisade wall, by contrast, occludes in both directions.

The first fruit concerns developments in conceptual foundations. That is, borrowing groundwork. Section four noted that modal epistemology and error statistics have different theoretical aims. Whereas modal epistemology typically and traditionally aims at characterising justified belief and knowledge, Mayo's severity conditions focus on test outcomes, especially in scientific practice, and whether inferences are warranted by overserved data. In what follows we highlight three significant differences that result from these different aims.[65]

Firstly, severe testing relates different relata from safety and sensitivity, at least according to their common formulations. Severe testing conditions connect a testing procedure, a particular body of data, and a hypothesis. They do not aim to describe the epistemic status of belief. Indeed many epistemologists of science argue that the assessment of belief is relatively unimportant, compared to other aims, for understanding the epistemic normativity of science. Staley and Cobb (2011, pp. 478–479) write, for example:

> [When recasting epistemology's internalism-externalism debate to better apply to statistical inference in scientific practice,] our first proposed modification requires a *shift from the appraisal of beliefs to the appraisal of assertions* as the proper object of epistemic evaluation. Whereas beliefs are private and individually held, at least in the paradigmatic cases, scientific knowledge is best regarded as a public and collective achievement. The activity of knowledge production in the sciences generally occurs within a *social structure and this involves acts of assertion* by scientists in various forums (i.e., preprints, publications, presentations, decisions taken in collaboration meetings, etc.). In fact, one could argue that it is intrinsic to scientific knowledge not merely that the acquisition of it often requires groups of people but that one aim of the scientific enterprise is a particular kind of *rationally persuasive communication* in which reasons are presented to other members of the community that will serve to underwrite, within that community, the status of particular claims as knowledge. [… We] are directing our attention to a distinct sense of scientific knowledge as *publicly accessible content* that arises from the socially organized efforts of individuals working in collaboration. (Emphasis added.)

This indicates that severe testing should not simply be recast as about belief. Scientific evidence is inherently socially distributed and its outputs might essentially involve communicative acts. Secondly, whether an individual's belief is justified depends on their *total* available evidence and epistemic resources. But the epistemic assessment of scientific inference and assertion might hinge on *restricted* bodies of information and community-approved inferential methods. Thirdly, severe testing conditions aim to *guide* inquiry, not merely assess its products. This includes steering scientific practices towards better methods of answering questions and away from faulty research practices, like those underlying the replication crisis.

These differences are significant and create challenges for the proposed unification. The two domains have different aims and subject matters. Mayo's 'guidance' aim leads her to focus on methods for auditing, for example, as an essential part of her full

[65] Cf. Staley and Cobb (2011) and Mayo (1997).

account. That is, she investigates how researchers should verify that their inferences are warranted. Safety and sensitivity, by contrast, are staunchly externalist conditions. One need not do anything to access or check whether they obtain.[66]

These differences are an obstacle to any straightforward unification of the two research programmes.[67] Yet they also create opportunities to draw on each other's developments. Social, applied epistemology increasingly foregrounds the epistemic practices of law, media, social media, education, and science communication. Epistemologists investigate how legal verdicts are warranted by evidence, for example, and when newspapers should report doubts about politicians' assertions.

These domains share pertinent features with scientific inquiry. This includes, for example, that questions of belief and knowledge are backgrounded relative to questions about warranted assertion, satisfying conventionalised epistemic benchmarks, communicating conclusions, publicly defending one's reasons and results, and those reasons being acceptable and intelligible to others. Permissible bodies of information and inference patterns might be restricted, either by convention, regulation, or necessity. Perhaps one's total evidence should not be used because the juror has background knowledge in a highly publicised trial. Questions of guidance and checking, including explicitly developing methods of inquiry and adjudication, are important in these domains.

Thus social epistemologists engaged in these emerging projects can adopt helpful groundwork from existing epistemology of science research. This includes ways to understand counterparts of belief, accessibility relations, the internalism–externalism distinction, available evidence, and epistemic position.[68]

We sketch one such example. Pritchard (2017) claims his safety condition can explain the epistemic normativity of legal proof. A common criticism holds that safety is too externalist to characterise legal proof.[69] An underlying reason for this critique is that formal legal findings must be publicly defensible and acceptable to various parties and accordingly factfinders should have some access to the reasons that secure the truth of their verdict. In response, Pritchard (forthcoming) introduced more internalist-friendly elements into his fuller account of appropriate legal verdicts. This includes the need for 'safeguards' and 'indications' that safety is satisfied. He writes, 'A *defensible* anti-risk strategy must thus *show that measures were taken* to ensure that the target risk event was modally far-off, such as by bringing in the kinds of *checks and balances* mentioned above' (Pritchard, forthcoming, pp. 3–4, emphasis added). Mere safety itself does not suffice; one must also assess whether the verdict is safe and explicitly take steps to insure it is.

[66] See Staley and Cobb (2011) for discussion. Mayo (1996) categorises common sources of error in scientific inference into groups. See also Mayo (2018, pp. 235–236).

[67] Our thanks to an anonymous reviewer for drawing our attention to this.

[68] Staley and Cobb (2011, especially §§2 and 5) exhibits this 'translatory' groundwork.

[69] Ebert, Durbach, and Smith (2020), Gardiner (2020, 2021b), Fratantonio (forthcoming). See Pritchard (forthcoming) for replies.

On Pritchard's resulting view, the externalist condition—safety—characterises what legal verdicts should aim at and helps guide methods of inquiry.[70] He writes, 'an information-relative assessment of risk is meaningfully *guided* by the modal account of risk, in that it offers the subject the means to assess, relative to their information regarding relevant features of the actual world, what the appropriate level of risk at issue is, and also what kinds of *strategies* would lower this risk' (Pritchard, forthcoming, pp. 4–5, emphasis added).

These substantial departures from the basic externalist condition find suggestive parallels in discussions of severe testing. Staley and Cobb (2011) describe how severe testing criteria provide externalist conditions that describe when hypotheses are supported by the evidence. They note that these conditions can guide how to develop research methods and check for sources of error, including especially in one's modelling assumptions. But a full account of why a particular inference is justified requires reference to that agent's 'epistemic situation' (Staley, 2012, pp. 22, 28–29). Staley and Cobb thus emphasise the need for both externalist and internalist elements in a full account of when statistical inference is justified by data.[71] And by satisfying the internalist conditions, the investigator acquires the ability to publicly articulate and defend their epistemic grounds for the inference.[72] This foreshadows Pritchard's recent emphasis on the ability to publicly defend legal verdicts.

These parallels merit further investigation. That is, perhaps when modal conditions are applied to social phenomena such as legal proof, the demands of the domain require augmenting the account with internalist-friendly conditions and existing parallel work in the epistemology of statistical inference can help guide the way.[73]

[70] See also Gardiner's (forthcoming-a) characterisation of the 'guiding' role of corroborative evidence. An existing body of evidence steers inquiry towards particular remaining uneliminated error possibilities and thereby towards investigating whether the initial evidence is misleading.

[71] See also Staley (2012, pp. 28–29). Mayo (2018, p. 236) endorses the resulting 'hybrid' interpretation.

[72] Staley and Cobb (2011, pp. 484–485) write,

> A thoroughgoing externalist, of course, would not accept our identification between the problem of justification and the question of one's ability to articulate supporting reasons, for on an externalist* account one can be justified in drawing conclusions even if one cannot access any reasons that support such a conclusion. [...].
>
> In reply, [recall] that our concern is with justification in the *socially situated contexts* of scientific inquiry and *communication*; it is the nature of these contexts, and not a prior commitment to internalism*, that grounds our understanding of the problem of justification. [Investigators] are responsible for *vindicating* their assertions and inferences in response to critical questioning from the community of investigators. In the absence of such a capacity for vindicating a conclusion, an investigator may be able to make statements that are objectively supported by evidence, but does not, thereby, contribute to the scientific pursuit of knowledge.' (Emphasis added. See also Staley (2012, p. 29)).

The parallels with legal practice merit investigation. In most jurisdictions jurors notably need not articulate their reasons underwriting their verdicts, which is an important difference. But a plausible epistemic demand on legal decision-making is that the reasons for verdicts are publicly articulatable. (Perhaps the lack of demand on jurors is prudential. It would be epistemically good if jurors conveyed their reasons, as scientists must. But lay jurors would perform poorly at this difficult task, which would undermine trust in the system. Scientists, by contrast, are trained at length in this skill.)

[73] Another suggestive parallel is that Pritchard's (forthcoming, pp. 4, 6–7) emphasis on 'information-relative assessment of risk' based on 'restricted bodies of evidence' and one's 'informational perspective'

Mayo pitches herself staunchly against Bayesianism, and offers an alternative view of the epistemic force of statistical data. On Mayo's view, merely assigning probabilities to hypotheses—even if those assignments accord with Bayesian updates based on evidence—is not enough because, under the Bayesian probabilist paradigm, there is no requirement that evidence must also be sensitive to error. This echoes the convictions of many mainstream non-formal epistemologists, who contend that probabilism cannot adequately capture whether and why a claim is warranted by the available evidence. Concordant reasons are offered: The evidence adduced in the Prisoner and Gendered Crime cases are inadequate for many purposes because it cannot address important possibilities of error, for example, and this requirement is interpreted subjunctively.

Colling and Szűcs (2018) argue that Mayo's approach and its significance testing kin 'find their strength where reasonable priors are difficult to obtain and when theories may not make any strong quantitative predictions' and 'exploratory contexts' in which inquirers simply want to know whether a phenomenon can be reliably measured. Bayesian approaches, by contrast, are better suited to adjudicating between rival quantitative models, or assigning credences or quantitative support for a claim. Colling and Szűcs advocate for a pragmatic pluralism. Rather than viewing Bayesian probabilism and Mayo's probativism as rivals, they suggest that each simply provides different methods that are appropriate in different contexts of inquiry. Regardless of whether their view is correct—we lack space to assess this here—their division of the terrain for each approach is revealing. In particular, it highlights a natural pairing of Mayo's error statistical probativism with mainstream epistemological theorising. From the perspective of mainstream non-formal epistemology, the former conditions characterise almost all contexts of inquiry, and the latter conditions are relatively marginal. Thus Mayo's approach has a natural home within orthodox, non-formal epistemological theorising. Mayo's severe testing provides an avenue for non-formal epistemologists to investigate the normative contours of statistical inference, reasoning from scientific data, and diagnosing and remedying flaws in scientific practice, including those highlighted by the replication crisis.

Mayo emphasises that statistical inferences are always initially made with reference to specific alternative hypotheses—not merely the whole cloth negation of the null hypothesis—and that inferences about those specific alternative hypotheses are only justified if they have passed severe tests.[74] Outright, non-comparative claims are justified when various potential sources of error, such as errors in the background

Footnote 73 continued
may echo, and be guided by, Staley and Cobb's (2011, p. 479) use of Achinstein's (2003, p. 20) 'epistemic situation' posit. On Staley and Cobb's view whether a hypothesis is in fact severely tested by the data is not relativised to a body of information and is a thoroughgoing externalist condition. But whether the scientific conclusion is justified depends on an epistemic situation and includes internalist elements. See also Staley (2012, pp. 28–29). Rather than simply referring to an individual's total available evidence, allowing for a socially-extended and conventionally-restricted evidence base better characterises inquiry in science and law.

[74] J. Neyman and E. S. Pearson were the first to modify Fisher's hypothesis testing to include reference to a class of possible alternative hypotheses (Lehmann, 1993; Fisher, 1925). Mayo's work extends Neyman and Pearson's hypothesis testing by quantifying the extent to which *specific* alternative hypotheses are severely tested, based on observed data (Neyman & Pearson, 1967). For details about Mayo's severe tests as extensions of Neyman-Pearson tests, see Mayo (2018, p. 142).

modelling assumptions, are ruled out.[75] These claims about testing specific alternatives are suggestively echoed by recent theorising about epistemic contrastivism and related relevant alternative theories.[76] Epistemic contrastivism claims that knowledge is not a binary relation between a subject and a proposition but a ternary relation between a subject, proposition, and a set of one or more (false) contrast propositions. Knowledge ascriptions, fully articulated, are not simply 'S knows that p' but rather 'S knows that p, rather than q'. But outright, non-contrastive knowledge ascriptions are nonetheless justified. The resulting view is not skeptical or error-theoretic about ordinary language practices of knowledge ascription. And—as with safety, sensitivity, and relevant alternatives conditions—contrastive conceptions might apply to epistemic phenomena other than knowledge. Accordingly one might compare epistemic contrastivism in mainstream epistemology with the comparativist structure of statistical inference to see whether they align, conflict, or offer mutual support.

There are fruitful parallels concerning epistemic value. Section one mentioned p-hacking methods that underwrite bad statistical inferences. Mayo diagnoses their flaws using her subjunctive severe testing condition: were the hypothesis H false, the data would nonetheless spuriously appear to support H. Rival 'performance-based' diagnoses, by contrast, appeal to long-run error rates: p-hacking is bad because it vitiates truth-to-falsity ratios in scientific inquiry. Mayo objects to this long-run performance-based diagnosis, noting the problem with p-hacking is not a matter of relative frequencies of erroneous inferences over time.[77] Instead inquirers care about truth in the particular case in hand. This better identifies problems with p-hacking: p-hacking diminishes the ability to avoid error in a particular case.[78]

This idea is echoed in objections to reliabilist theories of justification. Reliabilism holds that a belief is justified iff it is produced by a reliable cognitive belief-forming process.[79] We can sidestep the details; what matters here is that detractors claim reliabilism cannot explain epistemic value. They argue that what is valuable about a belief's being justified or known is not a matter of long-run performance. Instead what matters—the locus of value—is avoiding error and being assured in the particular case.

These objections to reliabilism tend to focus on the 'good case': reliably formed true belief. They claim reliability is only valuable insofar as it helps attain truth in a particular case, and so the value of being reliably formed is swamped by the value of the belief's being true. This objection holds that being reliably formed cannot add further value to a true belief.[80] But the objection to reliabilism is particularly sharp

[75] Staley (2012) and Mayo and Spanos (2004).

[76] Blaauw (2008) and Cockram and Morton (2017) survey contrastivism in epistemology. See also Schaffer (2005).

[77] Mayo (2018, p. 14).

[78] As noted above, there is often more than one flaw with such phenomena. P-hacking also undermines public trust in science, for example, and offers incentives that favour less forthright research.

[79] See Goldman and Beddor (2016) for a survey. Staley and Cobb (2011) and Woodward (2000) also find parallels between Mayo's error statistical approach and Goldman's reliabilism.

[80] See Zagzebski (2003, p. 13). Gardiner (2017) argues that safety conditions on knowledge generate a swamping problem because the core value is avoiding false belief, not modal distance from false belief. The value of sensitivity, by contrast, is arguably not similarly swamped. Sensitivity is valuable for, amongst other things, judgement's roles in appropriate action, assurance, and inference.

when—mirroring Mayo's focus on p-hacking—we shift attention from good cases to bad. The core problem with judgements formed through unreliable methods is not that in the long run such methods perform poorly. What matters is avoiding error in the case at hand. Inquirers desire accuracy on the particular occasion and error detection capacity is crucial for this. These dissatisfactions with reliabilism motivate shifting to safety and sensitivity accounts, with their emphasis on error detection capacities in the case at hand. This parallels Mayo's rejecting long-run performance-based accounts and favouring severe testing. Thus we see consilience between reliabilism's trouble explaining the epistemic value of knowledge and Mayo's criticisms of performance-based explanations of the disvalue of p-hacking.

These various parallels are worth highlighting even if ultimately one rejects Mayo's severe testing methods or the relevant alternatives framework. Indeed, appreciating the isomorphisms can aid detractors, since objections to one view might accordingly challenge the other. Perhaps reliabilism developed a rebuttal to the swamping problems that adherents of 'performance-based' views of statistical inference can repurpose, for instance.

Recall from the scales example that in order to infer a claim from observation, evidence must eliminate error possibilities. In normal cases, Ronda's weighing herself on one set of scales rules out all relevant error possibilities, and she can safely infer she weighs less than 117lbs. But further uneliminated error possibilities remain. This includes mundane (but nonetheless normally disregardably unlikely) error possibilities in which her scales are malfunctioning, more skeptical hypotheses, such as that her scales accurately weigh all objects except her, and the general rigged error possibility 'something other than H explains the observed results'. Evidence characteristically cannot eliminate all conceivable error possibilities and ruling out additional further error possibilities could be an endless task. This essay suggests safety can help characterise when to cease eliminating error possibilities.

Mayo notes there are practical reasons to cease inquiry. She argues that inquiries occur when we want to find things out and continuing to eliminate increasingly far-fetched error possibilities is a mistake when it thwarts epistemic and prudential goals. Continuing to eliminate further error possibilities for the claim that some infectious agents lack nucleic acid, for example, precludes learning about prion diseases, such as Alzheimer's.[81] These claims are familiar in the history of epistemology and arise in debates about inductive risk in the philosophy of science.[82]

Recently epistemology has turned towards addressing whether and why base rate evidence and other forms of 'merely numerical' evidence characteristically has less inquiry-closing potency than non-numerical evidence. (Recall the Prisoner and Gendered Crime examples, above.) Questions arise about the ethics and epistemology of failing to address morally distinctive error possibilities or sources of error. Existing

[81] See Mayo's discussion of prions and Kuru disease (Mayo, 2014, 2018, pp. 82–88; 108–110).

[82] Miller (2014) compares debates about pragmatic encroachment and inductive risk in science. See also Elliott and Richards (2017).

research in philosophy of science—both about inductive risk and statistical inference—can illuminate these debates. Conversely, recent insights in the ethics of belief can illuminate questions about inductive risk in science.[83]

The domains bring different strengths and priorities to questions about when to cease inquiry given inquiry costs and error risks. Philosophers of science contribute, amongst other things, an anchoring in concrete real-life examples and applicable formal models. And they aim towards carefully reflecting and guiding actual practices. Epistemologists offer orientation towards, for example, questions about closure and overall coherence of judgements. They also examine epistemic effects of social pressure and attention on whether inferences are justified. They ask, for example, whether commonly taking an error possibility seriously can itself render the possibility relevant, even if it is implausible or extremely unlikely to be true.

Finally, marrying these two domains illuminates the distinctive epistemic value of corroborating evidence.[84] Single source evidence can render a claim extremely probable, as lottery examples exemplify, but there is something distinctly compelling about independent or second-source evidence. Suppose a rape occurs. The perpetrator spiked a stranger's drink with the date rape drug Rohypnol and left DNA at the crime scene. A cold-hit DNA search—that is, trawling through DNA databases—identifies Jones as a leading suspect. This evidence makes it highly probable that Jones committed the crime. But the evidence does not address some error possibilities, such as those in which the forensics team framed Jones. For this illustration we can set aside questions about whether these error possibilities are relevant and so must be addressed. It depends on, amongst other things, whether such duplicity is normal and the judgement's purpose.

Suppose a second person, Corey, claims Jones purchased Rohypnol from him. The evidential force of this second piece of evidence is not fully captured by the increase in subjective probability of Jones's guilt. The probability given the available evidence does increase. But this increase cannot explain why the second piece of inculpatory evidence is so compelling. The probability was antecedently too high for the change to be so forceful. A change from 98 to 99% evidential probability does not register dramatically, for example. But Corey's corroborating testimony does.

The distinctive epistemic force of Corey's evidence is addressing error possibilities not addressed by the DNA cold-hit. Corey's testimony addresses many of the error possibilities in which Jones is innocent and the police framed him. It cannot eliminate them all. Given their divisible structure, remaining sub-possibilities are inevitable. But the only ones uneliminated by Corey's testimony are ones where Corey conspires

[83] See, for example, recent debates about the proof paradox, moral encroachment, and the epistemology of stereotyping. For recent work on pragmatic encroachment, see Kim and McGrath (2019). For relevant alternatives approaches to moral encroachment, see Bolinger (2020), Moss (2018a, 2018b, 2021), and Gardiner (2021b, §§6–7). Note that arguments from inductive risk in science don't support moral encroachment unless the property affected by moral value is the epistemic justification of belief or credence (Gardiner (ms)). But, as noted above, the relevant core of scientific practice might be best understood as acceptance, assertion, and other epistemic conduct, rather than belief. Gardiner (ms) explains why endorsing myriad connections between epistemic normativity and moral normativity is consistent with denying moral encroachment.

[84] Gardiner (forthcoming-a) builds on and expands these ideas about the epistemology of corroborating evidence.

with the police, has independent reason to lie, or Jones made the purchase but did not commit the rape and the cold-hit DNA match was extraordinarily bad luck.[85]

These remaining error sub-possibilities are notably more distant than original ones like the broad possibility that the police framed Jones. This underlies the epistemic power of Corey's testimony. The corroborating evidence guides future inquiry, furthermore, since investigators can proceed by addressing the possibility that Corey's testimony is part of a police conspiracy. Thus the dramatic shift in the landscape of uneliminated error possibilities explains the epistemic force of compelling corroborative evidence. This is far more notable than the increases in quantifiable evidential probabilities, which were antecedently too high to allow room for striking increases.

As incriminating evidence collects against a person, uncovering each new further piece of corroborating evidence can be increasingly compelling. That is, the epistemic force of each subsequent piece of evidence can increase as—and because—the inculpatory case grows. Each new piece can have a larger effect. This pattern is hard to explain if evidence's epistemic value is limited to increasing the claim's quantifiable probability. This is because the magnitude of each new increase in quantifiable probability will typically *decrease* as the inculpatory case grows. So the explanandum—the shift occasioned by accumulating corroborating evidence—can *increase* whilst the explanans—the magnitude of the probability increase—*decreases*, with each new piece of evidence. But this effect is predicted by the relevant alternatives (and severe testing) model: As uneliminated error possibilities are cumulatively chopped away, it becomes harder to maintain innocence. A conclusion is forced.

Similarly, Ronda's weighing herself on a second set of scales can, in many cases, settle the question in a way that does not simply amount to an increase in quantifiable evidential probabilities. The probability that her weight is less than 117lbs was already very high, given the results of the first scale. The second scale provides epistemic value not fully captured by the slight increase in the already very high probability. Results from the second scale address many of the closer error possibilities that were consistent with the first results, including many error possibilities in which the first scale was malfunctioning.[86]

We suggest the resources of severe testing and the relevant alternatives theory combine fruitfully to model the epistemic force of corroborating evidence. The relevant alternatives framework provides the epistemological structure for how error possibilities are ordered and how evidence can eliminate error possibilities. Mayo's research provides meticulous detail about how statistical reasoning and scientific methods eliminate those possibilities in practice.

[85] There are also error possibilities in which, for example, the putative victim knows about Jones's Rohypnol purchase, consented to sex, and (perhaps by subsequently dosing herself with Rohypnol to plant forensic evidence), framed Jones. But this error possibility is extremely farfetched, especially given—as noted above—the victim and Jones were strangers before the night in question and he was linked to the crime only through a cold-hit DNA match. Gardiner (2021a) describes how identifying uneliminated error possibilities can fuel undue doubt about rape accusations by creating an "ah-ha" feeling.

[86] Multiple accusations of crimes and replication studies in scientific practice have 'error possibility culling' epistemic value. They slice away remaining error possibilities, including deceit and researcher error. Gardiner (forthcoming-b) investigates 'possibility culling' and 'guiding' roles of corroborating evidence. Similar roles characterise the epistemic value of confirmation and 'robustness' in the philosophy of science (Soler, 2012). See also Staley (2008) and Mayo and Miller (2008).

Mayo celebrates Popper's emphasis on testing for sources of error. But she decries his approach—or lack thereof—to providing usable methods for detecting error. She writes,[87]

> [We must] erect a genuine account of learning from error—one that is far more aggressive than the Popperian detection of logical inconsistencies. Although Popper's work is full of exhortations to put hypotheses through the wringer, to make them "suffer in our stead in the struggle for the survival of the fittest" (Popper 1962, 52), the tests Popper sets out are white-glove affairs of logical analysis. If anomalies are approached with white gloves, it is little wonder that they seem to tell us only that there is an error somewhere and that they are silent about its source. *We have to become shrewd inquisitors of errors*, interact with them, simulate them (with models and computers), amplify them: *we have to learn to make them talk*.

Our proposal, then, follows this lead. Recent epistemological theorising has emphasised the importance of error sensitivity, understood as a subjunctive condition. Mayo advocates understanding statistical inference and scientific practice along the same lines. Insights from these two areas have remained largely segregated, which is a missed opportunity for both. It is time, we think, they talk.

Acknowledgements The core idea of this essay was presented at PhilStat, a summer seminar on the Philosophy of Statistics at Virginia Tech. We are grateful to the participants for their feedback, especially to Renee Jorgensen for particularly elucidative conversations throughout PhilStat. Thanks to Jelena Aleksic and Ben Martin for helpful insights about genetics and to Molly Kao and Elay Shech for their editorial excellence. Finally, many thanks to Jon Garthoff, Deborah Mayo, Duncan Pritchard, and two anonymous reviewers. Their comments greatly improved this essay.

References

Achinstein, P. (2003). *The book of evidence*. Oxford University Press.

Adams, F., & Clarke, M. (2005). Resurrecting the tracking theories. *Australasian Journal of Philosophy, 83*(2), 207–221.

Amaya, A. (2015). *Tapestry of reason*. Hart.

Arnholt, A. T. (2016). *PASWR2: Probability and statistics with R, Second Edition. R package version 1.0.2.* https://CRAN.R-project.org/package=PASWR2

Austin, J. (1946). Other minds. *Proceedings of the Aristotelian Society, 20*, 148–187.

Bandyopadhyay, P. S., Brittan, G., Jr., & Taper, M. L. (2016). *Belief, evidence, and uncertainty: Problems of epistemic inference*. Springer.

Blaauw, M. (2008). Contrastivism in epistemology. *Social Epistemology, 22*(3), 227–234.

Blome-Tillmann, M. (2015). Sensitivity, causality, and statistical evidence in courts of law. *Thought, 4*(2), 102–112.

Blome-Tillmann, M. (2017). "More likely than not" knowledge first and the role of bare statistical evidence in courts of law. In A. Carter, E. Gordon, & B. Jarvis (Eds.), *Approaches in epistemology and mind* (pp. 278–292). Oxford University Press.

Bolinger, R. J. (2020). The rational impermissibility of accepting (some) racial generalizations. *Synthese, 197*(6), 2415–2431.

Bradley, D. (2014). A relevant alternatives solution to the bootstrapping and self-knowledge problems. *Journal of Philosophy, 111*(7), 379–393.

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

[87] Mayo (1996, p. 4, emphasis added). See also Mayo (2018, p. 86).

Cockram, N., & Morton A. (2017). Contrastivism. *Oxford Bibliographies Online*.

Cohen, J. (1977). *The probable and the provable*. Oxford University Press.

Colling, L. J., & Szűcs, D. (2018). Statistical inference and the replication crisis. *Review of Philosophy and Psychology, 12*, 121.

Dang, H., & Bright, L. K. (2021). Scientific conclusions need not be accurate, justified, or believed by their authors. *Synthese* 1–17.

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

DeRose, K. (2017). *The appearance of ignorance: knowledge, skepticism, and context* (Vol. 2). Oxford University Press.

Dretske, F. (1970). Epistemic operators. *Journal of Philosophy, 67*(24), 1007–1023.

Dretske, F. (1971). Conclusive reasons. *Australasian Journal of Philosophy, 49*, 1–22.

Ebert, P., Durbach, I., & Smith, M. (2020). Varieties of risk. *Philosophy and Phenomenological Research, 101*, 432–455.

Elgin, C. Z. (2017). *True enough*. The MIT Press.

Elliott, K., & Richards, T. (2017). *Exploring inductive risk: Case studies of values in science*. OUP.

Enoch, D., Spectre, L., & Fisher, T. (2012). Statistical evidence, sensitivity, and the legal value of knowledge. *Philosophy and Public Affairs, 40*(3), 197–224.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Fleisher, W. (2020). Endorsement and assertion. *Noûs, 55*, 363.

Fratantonio, G. (forthcoming). Evidence, risk, and proof paradoxes: Pessimism about the epistemic project. *International Journal of Evidence and Proof*.

Gardiner, G. (2015a). Normalcy and the contents of philosophical judgements. *Inquiry, 58*(7), 700–740.

Gardiner, G. (2015b). Teleologies and the methodology of epistemology. In J. Greco & D. Henderson (Eds.), *Epistemic evaluation: Purposeful epistemology* (pp. 31–45). Oxford University Press.

Gardiner, G. (2017). Safety's swamp: against the value of modal stability. *American Philosophical Quarterly, 54*(2), 119–129.

Gardiner, G. (2018). Legal burdens of proof and statistical evidence. In D. Coady & J. Chase (Eds.), *Routledge handbook of applied epistemology* (pp. 171–195). Routledge.

Gardiner, G. (2019a). Legal epistemology. In D. Pritchard (Ed.), *Oxford bibliographies: Philosophy*. OUP.

Gardiner, G. (2019b). The reasonable and the relevant: Legal standards of proof. *Philosophy & Public Affairs, 47*(3), 288–318.

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

Gardiner, G. (2021a). Banal skepticism and the errors of doubt: On ephecticism about rape accusations. *Midwest Studies in Philosophy, 45*, 393–421.

Gardiner, G. (2021b). Relevance and risk: Relevant alternatives and the epistemology of risk. *Synthese, 199*, 481–511.

Gardiner, G. (forthcoming-a). Corroboration. *American Philosophical Quarterly.*

Gardiner, G. (forthcoming-b). Legal evidence and knowledge. In M. Lasonen-Aarnio & C. Littlejohn (Eds.), *Routledge handbook of the philosophy of evidence*. Routledge.

Gardiner, G. (forthcoming-c). Pragmatism, skepticism, and over-compatibilism: On Michael Hannon's *What's the point of knowledge?'. Inquiry.*

Gardiner, G. (ms). Against the new ethics of belief: The morass of moral encroachment and doxastic partiality.

Gelman, A., Haig, B., Hennig, C., Owen, A., Cousins, R., Young, S., Robert, C., Yanofsky, C., Wagenmakers, E.J., Kenett, R., & Lakeland, D. (2019). Many perspectives on Deborah Mayo's. In *Statistical inference as severe testing: How to get beyond the statistics wars*. Cornell University Statistics.

Glymour, C. (1980). *Theory and evidence*. Princeton University Press.

Goldman, A. (1976). Discrimination and perceptual knowledge. *Journal of Philosophy, 73*, 771–791.

Goldman, A., & Beddor, B. (2016). Reliabilist epistemology. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*.

Hannon, M. (2015). The universal core of knowledge. *Synthese, 192*(3), 769–786.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of P-hacking in science. *PLoS Biology, 13*(3), e1002106.

Heller, M. (1989). Relevant alternatives. *Philosophical Studies, 55*(1), 23–40.

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

Ho, H. L. (2008). *A philosophy of evidence law*. Oxford University Press.

Ichikawa, J. J. (2017). *Contextualising knowledge*. Oxford University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.

Jackson, E. (2018). Belief, credence, and evidence. *Synthese, 197*, 5073.

Kim, & McGrath, M. (Eds.). (2019B). *Pragmatic encroachment in epistemology*. Routledge.

Lawlor, K. (2013). *Assurance*. Oxford University Press.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88*(424), 1242–1249.

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

Mayo, D. (1996). *Error and the growth of knowledge*. University of Chicago.

Mayo, D. (1997). Duhem's problem, the Bayesian way, and error statistics, or "what's belief got to do with it?" *Philosophy of Science, 64*(2), 222–244.

Mayo, D. (2005). Evidence as passing severe tests: Highly probable versus highly probed hypotheses. In P. Achinstein (Ed.), *Scientific evidence* (pp. 95–127). Johns Hopkins.

Mayo, D. (2014). Learning from error: How experiment gets a life (of its own). In M. Boumans, G. Hon, & A. Petersen (Eds.), *Error and uncertainty in scientific practice* (pp. 57–77). Pickering and Chatto.

Mayo, D. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

Mayo, D., & Miller, J. (2008). The error statistical philosopher as normative naturalist. *Synthese, 163*(3), 305–314.

Mayo, D., & Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science, 71*, 1007–1025.

Mayo, D., & Spanos, A. (2011). Error statistics. In *Handbook of philosophy of science, Volume 7 Philosophy of statistics*, (General editors: Dov M. Gabbay, Paul Thagard and John Woods; Volume eds. Prasanta S. Bandyopadhyay and Malcolm R. Forster.) Elsevier.

Mayo-Wilson, C. (2018). Epistemic closure in science. *Philosophical Review, 127*(1), 73–114.

McKinnon, R. (2013). Lotteries, knowledge, and irrelevant alternatives. *Dialogue, 52*(3), 523–549.

Melchior, G. (2019). *Knowing and checking: An epistemological investigation*. Routledge.

Melchior, G. (2020). Sensitivity principle in epistemology. In *Oxford bibliographies online.*

Melchior, G. (forthcoming). A modal theory of discrimination. *Synthese.*

Miller, B. (2014). Science, values, and pragmatic encroachment on knowledge. *European Journal for Philosophy of Science, 4*(2), 253–270.

Moss, S. (2018a). Moral encroachment. *Proceedings of the Aristotelian Society, 118*(2), 177–205.

Moss, S. (2018b). *Probabilistic knowledge*. Oxford University Press.

Moss, S. (2021). Knowledge and legal proof. In *Oxford studies in epistemology* (Vol. 7). Oxford University Press.

Neyman, J., & Pearson, E. (1967). *Joint statistical papers*. University of California Press.

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

Palmira, M. (2020). Inquiry and the doxastic attitudes. *Synthese, 197*, 4947–4973.

Pardo, M. S. (2018). Safety vs. sensitivity: Possible worlds and the law of evidence. *Legal Theory, 24*(1), 50–75.

Pinillos, Á. (forthcoming). Bayesian sensitivity principles for evidence based knowledge. *Philosophical Studies.*

Pritchard, D. (2002). Recent work on radical skepticism. *American Philosophical Quarterly, 39*, 215–257.

Pritchard, D. (2005). *Epistemic luck*. Oxford University Press.

Pritchard, D. (2007). Anti-luck epistemology. *Synthese, 158*, 277–297.

Pritchard, D. (2009). Safety-based epistemology: Whither now? *Journal of Philosophical Research, 34*, 33–45.

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

Pritchard, D. (2015). Risk. *Metaphilosophy, 46*, 436–461.

Pritchard, D. (2017). Legal risk, legal evidence and the arithmetic of criminal justice. *Jurisprudence, 9*(1), 108–119.

Pritchard, D. (forthcoming). In defence of the modal account of legal risk. *Synthese*.

Rabinowitz, D. (2014). The safety condition for knowledge. In *Internet encyclopedia of philosophy*.

Redmayne, M. (2008). Exploring the proof paradoxes. *Legal Theory, 14*, 281–309.

Ross, L. (2021). Recent work on the proof paradox. *Philosophy Compass., 15*(6), e12667.

Roush, S. (2007). *Tracking truth: Knowledge, evidence, and science*. Oxford University Press.

Russell, B. (1948). *Human knowledge: Its scope and its limits*. Allen & Unwin.

Rysiew, P. (2006). Motivating the relevant alternatives approach. *Canadian Journal of Philosophy, 36*(2), 259–279.

Schaffer, J. (2005). Contrastive knowledge. *Oxford Studies in Epistemology, 1*, 235–271.

Soler, L. (Ed.). (2012). *Characterizing the robustness of science: After the practice turn in philosophy of science*. Springer.

Sosa, E. (1999). How to defeat opposition to moore. *Philosophical Perspectives, 13*, 141–154.

Spanos, A. (1999). *Probability theory and statistical inference*. Cambridge UP.

Staley, K. (2008). Error-statistical elimination of alternative hypotheses. *Synthese, 163*, 397–408.

Staley, K. (2012). Strategies for securing evidence through model criticism. *European Journal for Philosophy of Science, 2*, 21–43.

Staley, K. (ms). Two ways to rule out error: Severity and security.

Staley, K., & Cobb, A. (2011). Internalist and externalist aspects of justification in scientific inquiry. *Synthese, 182*, 475–492.

Stine, G. (1976). Skepticism, relevant alternatives, and deductive closure. *Philosophical Studies, 29*, 249–261.

Williamson, T. (2000). *Knowledge and its limits*. OUP.

Williamson, T. (2007). *The philosophy of philosophy*. Blackwell.

Woodward, J. (2000). Data, phenomena, and reliability. *Philosophy of Science, 67*, S163–S179.

Zagzebski, L. (2003). The search for the source of the epistemic good. *Metaphilosophy, 34*, 12–28.