Epistemic Closure in Science

Conor Mayo-Wilson

Abstract

Epistemic closure (EC) is the thesis that knowledge is closed under known entailment. Although several theories of knowledge violate EC, failures of EC seem rare in science. I argue that, surprisingly, there are genuine violations of EC according to theories of knowledge widely endorsed in the sciences.

Introduction

Epistemic closure (EC) is the thesis that knowledge is closed under known entailment, or in other words, if one knows that φ and one knows that φ entails ψ, then one knows that ψ.[1] Many counterfactual and "relevant alternatives" theories of knowledge violate EC.[2] Nonetheless, failures of EC in science seem rare, if not impossible. For example, it is known that the mass of a proton is approximately $1.67 \cdot 10^{-27}$ kg, whereas that of an electron is $9.11 \cdot 10^{-31}$ kg. So it would be bizarre if, according to standards for scientific knowledge, physicists could not know that a proton is more massive than an electron on the basis of deduction.

The lack of scientific counterexamples to EC suggests three possibilities. First, philosophical theories of knowledge that violate EC might diverge from

[1] It is difficult to state a plausible, precise version of EC. EC is generally defined as a weaker conditional, where the antecedent requires that one believes ψ on the basis of inferring it from φ, that one has no defeaters for ψ, and more. My counterexamples to EC also violate these weaker conditionals. Moreover, many philosophers distinguish between single-premise and multi-premise closure. The former asserts that knowledge is closed under known entailment involving a rule of inference with exactly one premise, whereas the latter allows for rules with multiple premises. I consider both theses below.

[2] To my knowledge Dretske [1970] first rejected EC. Nozick [1981] made the rejection a centerpiece of his theory of knowledge, but Roush [2010] denies that Nozick's view violates EC in ways that are typically thought. Sosa [1999] claims that safe belief is not closed under entailment, but fails to provide any examples. Kvanvig [2004], Murphy [2006], and Alspector-Kelly [2011] each provide examples. In contrast, Luper [2012] argues that safe belief, suitably formalized, satisfies EC. Holliday [2015] argues that, when externalist theories of knowledge are formalized in epistemic logic, failures of EC are ubiquitous.

the standards that scientists use to judge what constitutes knowledge, and the latter standards might satisfy EC. Alternatively, failures of EC might be confined to inferences about skeptical hypotheses that rarely arise in science. Finally, perhaps violations of EC in science have been overlooked. I argue for the third option.

Section one sketches two theories of scientific knowledge, which I call the CCS and SP theories respectively.[3] These theories, I conjecture, are widely-accepted in the sciences, if only implicitly.[4] But even if my conjecture is wrong because CCS and SP characterize justified belief or evidence rather than knowledge, it would still be interesting that justified belief or evidence is not closed under implication in common, quantitative scientific contexts.

Section two explains two new theorems.[5] The first asserts that a hypothesis is knowable in principle according to CCS if and only if it is knowable according to a theory that identifies knowledge with true belief satisfying a probabilistic version of the safety principle. The second asserts that if a hypothesis is SP knowable, then it is knowable according to a theory that identifies knowledge with true belief satisfying probabilistic analogs of Nozick's tracking conditions. Hence, theories of scientific knowledge resemble philosophical theories that violate EC, pace the first option above.

In Section three, I argue that, contra the second option, when scientific hypotheses are inferred from the outcomes of several statistical tests, EC is routinely violated by SP and the two theories of knowledge identifying knowledge with true belief satisfying probabilistic analogs of (i) the safety

[3] CCS stands for "confident, consistent, and stable" and SP stands for "size and power."

[4] According to [Gigerenzer, 1990, p. 210-211], "the dream of mechanization of knowledge" drove psychologists' adoption of a hybrid of Fisherian and Neyman-Pearsonian testing in the twentieth century. In practice, many significance tests in psychology have lower power [Cohen, 1988, Sedlmeier and Gigerenzer, 1989], which leads [Schmidt, 1996, p. 115] to conclude that the "reliance on significance testing retards the development of *cumulative knowledge*" [my emphasis]. Gigerenzer and Schmidt's criticisms indicate that psychologists use statistical tests to acquire knowledge, and moreover, that high power is a necessary criterion for knowledge just as SP requires. In medical science, [du Prel et al., 2010, p. 343] claim "knowledge is increasingly based on empirical studies and the results of these are usually presented and analyzed with statistical methods." du Prel and colleagues then review nearly 2000 articles in medical research, and they argue that only a handful of statistical tests are used and that such tests are considered to be reliable only if they have low size and power of at least 80%. Although SP may be held only implicitly in the sciences, similar theories are explicitly defended by some philosophers. For instance, see Mayo [1996]. Further, [Giere, 1975] more-or-less defends SP when he writes, "'An alternative to a full-fledged decision theoretic approach is to take 'accepting H' to mean something like *adding H, tentatively, to the body of scientific knowledge*" [my emphasis]. Moreover, [Giere, 1976] advocates the Neyman-Pearsonian methodology inspiring SP.

[5] Proofs of theorems appear in the technical appendix.

condition and (ii) Nozick's tracking conditions respectively.

Section four explains why scientific counterexamples to EC seem rare even though some theories of scientific knowledge violate EC. There, I argue that if one's beliefs are based on confidence intervals in the right way, then knowledge is closed under known entailment according to the all of the theories I consider. Finally, in section five, I summarize the broader importance of my technical results, which I discuss briefly now.

Because scientific knowledge plays a crucial role in political decision-making, assessments of legal responsibility and more, violations of EC in science have important moral and political consequences that typically-discussed violations of EC (e.g., concerning external-world skepticism) do not. For example, suppose drug manufacturers were required to disclose the results of preliminary research only if they knew a drug had "some negative side effect." After all, legislators could not be expected to anticipate every possible harm a drug might cause. Nonetheless, cigarette manufacturers would not be blameless if, according to widely-accepted standards of knowledge, they knew only that cigarettes cause lung cancer and not that cigarettes cause "some" harm. Rather, the general public and legal experts would recommend revising standards for scientific knowledge to accommodate established practices for assigning blame and legal responsibility. In particular, legal experts would argue that anyone who knows "cigarettes cause cancer" also knows "cigarettes cause harm." The violations of EC that I discuss have the same logical form as this example, and so a central consequence of my arguments is that we ought to revise some implicit, but common standards for scientific knowledge.

1 Classical Statistics in Science

Scientific knowledge often relies on statistics. For example, we know the charge of the electron to thirty decimal places. How? Physicists have made repeated measurements with precise instruments, and they use the best statistical procedures to analyze the resulting data.

Of course, statistical analysis is rarely sufficient for knowledge. If we had no reason to believe that electrons exist or that all electrons share the same charge, then estimates of "the" charge of an electron would not produce knowledge, even if they were accurate. In general, it is necessary to justify the assumptions of statistical procedures in order for their products to count as knowledge. Often, some well-established scientific theory (e.g., electrostatics) justifies the underlying statistical model.

Moreover, statistical estimation is sometimes unnecessary for knowledge. For example, in the first clinical trials in which citrus fruit was compared with other treatments for scurvy (e.g., drinking sea water), scientists lacked even rudimentary statistical techniques for comparing the various treatment groups. Nonetheless, those experiments produced knowledge that eating by citrus fruit can cure scurvy.

So the use of statistics is neither necessary nor sufficient for scientific knowledge. Nonetheless, there are many scientific facts about the microscopic realm, the distant past, and the faraway recesses of the universe that are known primarily because of the existence of large data sets and rigorous statistical estimation procedures. If we want to understand how scientific knowledge is acquired in these circumstances, therefore, we must learn what features our "best" statistical estimation procedures possess.

To discuss "good" statistical estimators, I use the following running example. Suppose a medical researcher is interested in the one-year mortality rate of women treated with a new breast cancer drug. Define the *efficacy* of the drug to be one minus the one-year mortality rate of the patients treated with it. The drug is intended for women with stage-two breast cancer, for which the current (one-year) mortality rate is 7%. Hence, the researcher is interested in whether the drug's efficacy exceeds 93%. She conducts a clinical trial, and then she estimates the drug's efficacy using some statistical procedure.

What criteria are used to evaluate the reliability of the researcher's procedure? It depends upon the type of estimate she provides. The researcher might provide a *point estimate* (i.e., a single number) that represents her "best guess" of the drug's efficacy. Alternatively, she might provide an *interval estimate*, which is a range of values. Or if she is interested only in whether the efficacy exceeds a particular value (e.g., that of the conventional treatment), her "estimate" might only be an answer to the question, "Should I accept the hypothesis that the old treatment is at least as effective as the new?" *Hypothesis tests* are used to answer this last question.

In the next two subsections, I discuss four criteria from classical statistics that characterize different senses in which point estimators, interval estimators, and hypothesis tests might be called reliable. Although there are many different criteria statisticians use to characterize reliability, I focus on size, power, confidence, and consistency because (generalizations of) these criteria can be applied to both hypothesis tests and estimators, whereas some criteria (e.g., unbiasedness) can be applied only to point estimators or in-

terval estimators.[6]

1.1 Interval and Point Estimation

Confidence intervals are ubiquitous in science. What distinguishes confidence intervals from other interval estimates?

Consider first what makes an interval estimate erroneous. For example, if the medical researcher estimates the drug's efficacy is between 95% and 96%, and if the efficacy is actually 92%, then her estimate is wrong. In general, an interval estimate (of the efficacy of a drug, charge of an electron, etc.) is wrong precisely if it does not contain the true value of the quantity.

In my example, the medical researcher wants to minimize "the" probability that her estimate fails to contain the true efficacy. But that requirement is imprecise. If the efficacy is actually 94% and if the researcher would conjecture that the rate is between 93% and 95%, regardless of her data, then there is zero chance that her estimate is wrong in the actual world. However, her estimate is wrong with probability one in every world in which the efficacy is greater than 95%. Thus, there is no single probability of error; there are many. Because the researcher does not know the drug's efficacy, she desires an estimator that has a low chance of error *regardless of the efficacy's true value*. That's what high confidence guarantees.

Formally, let $[0, 1]$ be the set of numbers between zero and one inclusive, which represent the possible values of the drug's efficacy. Suppose the researcher's estimator is represented by a sequence of functions $\langle \hat{e}_n \rangle_{n \in \mathbb{N}}$. For each natural number $n = 1, 2, \ldots$, the function $\hat{e}_n$ takes n data points as input (i.e., a list of which of the first n patients survives for a year), and it outputs an interval of plausible mortality rates. Given any possible mortality rate $\theta \in [0, 1]$ and any natural number n, let $P_\theta(\theta \in \hat{e}_n)$ denote the probability that, after observing n patients, the medical researcher's estimate contains θ *if the efficacy is equal to* θ. Let $a > 0$ be some small probability. The medical researcher's estimate has confidence $1 - a$ at sample size n if

[6] See Appendix C for an explanation of how to extend the concepts of size and power to point and interval estimators.

$P_\theta(\theta \in \hat{e}_n) \geq 1 - a$ for all $\theta \in [0, 1]$, or equivalently if[7]

$$\textsc{Confidence}(\hat{e}_n, a) : \inf_{\theta \in [0,1]} P_\theta(\theta \in \hat{e}_n) \geq 1 - a.$$

Scientists often desire methods with 95% confidence (i.e., $a = .05$), but researchers may choose a larger or smaller a depending upon the context.

High confidence, however, is not sufficient. Here is a 100% confidence interval: guess that the efficacy is between zero and one inclusive. A researcher who forms her belief in this way will always conjecture an interval that contains the true value. Yet she does so at the expense of believing only tautologies. Scientists have dozens of recipes for balancing informativeness against error. I discuss one property, called "statistical consistency" (or consistency, simpliciter), that can be used to navigate this trade-off.

Imagine the medical researcher is asked for her "best guess" (i.e., a point estimate) of the drug's efficacy. For simplicity, assume her estimate equals the fraction of patients who die during the one-year trial. This fraction is called the *sample mean*. Of course, the researcher's guess is probably not exactly right. Moreover, if she treats only a few patients, her guess might be far off. For example, if she treats exactly ten patients, it is possible that all ten patients die, even if the drug's efficacy is really 99%. Luckily, it can be shown that, because she uses the sample mean estimator, the researcher's estimate increases in accuracy as her sample size grows: with increasingly high probability, her "best guess" of the efficacy will approach the true value. This fact is called statistical consistency.

The concept of consistency can also be applied to interval estimators. How? Call an interval estimator "consistent" if the endpoints of its intervals approach the true efficacy. Some standard confidence intervals are consistent in this sense.

Formally, for every possible efficacy $\theta \in [0, 1]$ and for any positive number $r > 0$, let $B_r(\theta)$ denote the set of numbers that differ from θ by less than r, i.e., $B_r(\theta) = \{\upsilon \in [0, 1] : |\upsilon - \theta| < r\}$. Let $P_\theta(\hat{e}_n \subseteq B_r(\theta))$ denote the probability that, if the efficacy were θ and the researcher observed n patients, all of the numbers in her estimate would differ from θ by less than

[7]Here, "inf" abbreviates "infimum," which is like a minimum. The set of numbers $(0, 1)$ between 0 and 1 (exclusive) has no minimum because for any small number (e.g., .01), there is another number (e.g., .001) that is smaller but still strictly greater than zero. However, $(0, 1)$ does have an infimum, namely, zero. An infimum is a greatest lower bound. Similarly, "sup" means "supremum," which is a least upper bound. The supremum of $(0, 1)$ is 1.

r. The researcher's estimator is called *consistent* if

$$\textsc{Consistent}(\hat{e}): \lim_{n\to\infty} P_\theta(\hat{e}_n \subseteq B_r(\theta)) = 1 \text{ for all } \theta \in [0,1] \text{ and all } r > 0.$$

Consistency is a "large sample" property. It says that, as the medical researcher acquires increasing amounts of data, her estimates become more accurate. There are also "large sample" notions of confidence. Let $a > 0$ be some small probability, and say the researcher's estimator has *asymptotic confidence* $1 - a$ if, as her sample becomes large, the minimum (formally infimum) chance that her estimate contains the true efficacy becomes at least $1 - a$, i.e., if $\lim_{n\to\infty} \inf_{\theta\in[0,1]} P_\theta(\theta \in \hat{e}_n) \geq 1 - a$. So much for point and interval estimators. How are hypothesis tests evaluated?

1.2 Hypothesis Tests

Suppose the medical researcher must decide whether the new treatment is more effective than the old. She might assume, by default, that the new drug is no better, i.e., that the new drug's efficacy is no more than 93%; call this the *null hypothesis*. But if the data are sufficiently strong, she might reject the null hypothesis and believe the *alternative hypothesis* that the efficacy is greater than 93%. Formally, the null hypothesis can be represented by the set $\Theta_0 = [.93, 1]$ of numbers (i.e., efficacies) between .93 and 1. Similarly, the alternative may be represented by the set $\Theta_1 = [0, .93)$ of numbers that are at least zero but strictly less than .93.

Hypothesis tests are used to decide between competing statistical hypotheses. What makes a hypothesis test good? Statisticians often distinguish between Type I and Type II errors. A researcher commits a Type I error when she rejects a true null hypothesis. The probability of a Type I error varies with the drug's efficacy. For example, if the drug in the running example is lethal (e.g., its efficacy is near zero), then there is nearly zero chance that the researcher will falsely reject the null hypothesis, which recall states the efficacy is at least 93%. However, if the true efficacy is precisely 93%, then the chance of a Type I error is higher, as the researcher might observe an unusually healthy sample of patients who survive at rates higher than the population average. The maximum (technically, supremum) chance of a Type I error is called the *size* of a test.

Formally, let $\hat{e} = \langle \hat{e}_n \rangle_{n\in\mathbb{N}}$ be a sequence of functions that represent the researcher's hypothesis test, i.e., $\hat{e}_n$ takes n data points as input and outputs either the null hypothesis Θ_0 or the alternative Θ_1.[8] Then let

[8] Statisticians typically define a hypothesis test to be a procedure for determining only

$P_\theta(\Theta_0 \subseteq \hat{e})$ represent the probability that the researcher does not reject the null hypothesis, if the efficacy is θ. Then the size of the researcher's test is defined to be $\sup_{\theta \in \Theta_0} P_\theta(\Theta_0 \subseteq \hat{e})$.

A researcher commits a Type II error when the alternative hypothesis is true and yet she retains the null. Like Type I error, the chance of a Type II error depends upon the drug's efficacy. If the drug's efficacy is zero, the researcher will likely correctly reject the null hypothesis, but if it is 92.9%, then the chance of a Type II error is high.

Formally, for each efficacy θ, let $P_\theta(\Theta_0 \subseteq \hat{e}_n)$ denote the probability that the researcher fails to reject the null Θ_0 (after treating n patients). If the null hypothesis is false, then ideally $P_\theta(\Theta_0 \subseteq \hat{e})$ will be small, i.e., $1 - P_\theta(\Theta_0 \subseteq \hat{e})$ will be high. The *power function* of the researcher's test takes as input a possible efficacy θ for which the null hypothesis is false, and it outputs the probability $1 - P_\theta(\Theta_0 \subseteq \hat{e})$ of correctly rejecting the null. A test is considered "powerful" if the power function is close to one for all θ.

Classical statisticians recommend tests with low size and high power, but unfortunately, minimizing Type I and Type II error simultaneously is impossible. One can avoid Type I error by always accepting the null hypothesis, and one can avoid Type II error by always rejecting the null. Thus, many statisticians recommend fixing one's maximal tolerance for Type I error (e.g., tolerating no more than 5% chance of Type I error) and then finding a test that maximizes power subject to that constraint.[9]

Hypothesis tests have attracted increasing scrutiny in recent years, and in some sciences, they are being replaced by confidence intervals. Nonetheless, significance testing remains entrenched in some fields. In particle physics, for example, the existence of a new particle is justified by hypothesis

whether one ought to reject the null hypothesis. I have followed this convention in the text. But this restriction is useful only if one is faced with a binary choice to reject the null or not, and it leads to needless verbal acrobatics otherwise. For example, classical statisticians routinely emphasize that evidence against a null hypothesis is not evidence for a particular alternative, and vice versa. They also adopt terms like "retaining" the null hypothesis to avoid recommending "acceptance." Rather than engage in verbal acrobatics, I define Ans be the result of closing $\{\Theta_0, \Theta_1\}$ under unions and complements minus the empty set, i.e., $\text{Ans} = \{\Theta_0, \Theta_1, \Theta_0 \cup \Theta_1, \Theta_0 \cup \neg\Theta_1, \neg\Theta_0 \cup \Theta_1, \neg\Theta_0 \cap \neg\Theta_1, \Theta\}$. Note Θ_1 need not be the complement of Θ_0. I define an **hypothesis test** to be an estimator $\hat{e}$ such that $\hat{e}_n(x) \in \text{Ans}$ for all data sequences x. Call a hypothesis test **binary** if $\hat{e}(x) \in \{\Theta_0, \Theta_1\}$ for all x. All the theorems in the text are about binary tests, but my definition allows one to represent suspension of judgment, rejection of both the null and alternative, rejection of the null without accepting the alternative, and more.

[9]Importantly, sometimes no test is powerful. Nonetheless, classical statisticians often look for *uniformly most powerful* tests, which maximize power for all $\theta \in \Theta_1$.

tests satisfying a very stringent requirement called the *5σ criterion.*[10]

Statistical procedures vary substantially, and to list all criteria that are used to evaluate estimators and hypothesis tests would be impossible. But confident, consistent estimators are often considered the ideal in interval and point estimation, and tests with high power and low size are often considered ideal in hypothesis testing. So if scientific knowledge is sometimes acquired from our best statistical procedures, and if our best statistical procedures are consistent and confident (or have low size and high power), then it is fruitful to investigate a theory in which scientific knowledge is identified with true belief acquired from consistent and confident estimators (or from tests with low size and high power).[11] This is the goal of the next section.

2 Scientific Knowledge

To develop theories of scientific knowledge, I first discuss how scientists use statistical procedures to form beliefs. Consider the running example again. The medical researcher might believe the drug's efficacy is a specific value, say 94%. Alternatively, she might believe only that the efficacy is between 93% and 94%. In both cases, her beliefs can be represented by a set of numbers. In the first case, her belief is representable by the singleton set $\{.94\}$ and in the second by the interval $[.93, .94]$.

In general, I represent propositions by sets of numbers, and I identify the researcher's beliefs with a set of propositions. For simplicity, I assume that the researcher has a logically strongest belief E, in the sense that she believes all propositions (about the drug's efficacy) that follow from E, but she believes nothing else.[12] I assume she disbelieves any proposition that is the negation of one of her beliefs, and she suspends judgment about all else. For example, suppose her strongest belief is that the drug's efficacy is between 93% and 95%. Then she believes that the drug's efficacy is at least 92%; she suspends judgment about whether the rate is between 94% and 96%, and she disbelieves that the rate is less than 93%.

Formally, if the researcher's strongest belief is E, then she believes F if $E \subseteq F$; she disbelieves F if $E \subseteq \neg F$, where $\neg F = [0,1] \setminus F$ is the set of numbers in $[0,1]$ that are not in F. She suspends belief in F otherwise.

[10] See [Franklin, 2014] for a history of the criterion.

[11] Again, I stress these theories of knowledge are applicable only in certain domains, as there are cases of scientific knowledge in which no statistical reasoning is used.

[12] For the rest of the paper, I use the capital letters E, F and G to denote propositions. For brevity, I ignore the distinction between a proposition and its representation, so that I often speak of a proposition E as if it were just a set of numbers between zero and one.

I assume, therefore, that the researcher is *logically omnicredulous*, i.e., her beliefs about the drug's efficacy are closed under logical consequence.[13] This assumption is important because, otherwise, the researcher may know E and $E \to F$[14] but fail to know F because she doesn't believe F. By assuming she is omnicredulous, we can determine which violations of EC are due to conditions necessary for knowledge rather than the researcher's limited logical abilities.[15]

How does the scientist form her strongest belief? For simplicity, I assume that she believes the output of her estimator or test. So if her hypothesis test recommends rejecting the null hypothesis, then she believes the null hypothesis is false. Alternatively, she might believe all propositions that follow from a confidence interval.

Formally, again let $\hat{e} = \langle \hat{e}_n \rangle_{n \in \mathbb{N}}$ represent the researcher's estimator or hypothesis test, where $\hat{e}_n$ takes n many data points as input and then outputs a proposition, which will be the researcher's strongest belief. If the researcher uses a point estimator, then the output of $\hat{e}_n$ will be a proposition of the form $\{\theta\}$, i.e., a proposition containing a single number. Interval estimators output interval propositions; hypothesis tests output either the null or alternative hypothesis under investigation. If the drug's efficacy is θ and E is some proposition, then $P_\theta(\hat{e}_n \subseteq E)$ therefore represents the probability that, in a world in which the efficacy is θ, the researcher believes E after observing n patients.

Importantly, if possible worlds are not representable by numbers or functions, then many common statistical "estimators" will not be estimators in my sense. Why? Statistical estimators output numbers or functions; my estimators output propositions. For instance, suppose a medical researcher wants to estimate the number of treated patients who will survive the next decade. Then some statistical procedures might recommend estimating the number is between 92.3 and 92.6 even though it's impossible that a fractional number of people have survived. But estimators, in my sense, output sets of possible worlds, and if only discrete values of some quantity represent possible worlds, then only sets of discrete quantities will be output by what

[13] If $E \subseteq F$, it does not follow that E *logically* entails F (e.g., if E asserts "apples are red" and F asserts "apples are colored"), and so the general version of my assumption is stronger than *logical* omnicredulity.

[14] $E \to F$ abbreviates the set $\neg E \cup F$.

[15] I also assume one believes, disbelieves, or suspends judgment about a proposition. In contrast, philosophers, social scientists, and statisticians often model individuals as having degrees of belief, which are often represented by probabilities. In order to discuss EC, however, it is most natural to model an agent as having full beliefs; it is not clear what the analog of EC is when agents' beliefs are represented by probability functions.

I call an "estimator." Therefore, when worlds are not numerically representable, it's better to think of estimators, in my sense, as the encoding a two-stage process. First, a reliable statistical estimator is employed, and second, one uses the statistical estimate and background scientific theory to generate one's beliefs.

Even when worlds are numerically representable, not all common statistical estimators are reasonable belief-forming methods. For example, it would be unreasonable to believe that the efficacy of a drug is *exactly* 92.7%, but that's what the medical researcher would believe if she used a point estimator in the way I describe above. If worlds are representable by some *continuous* quantity, like in the running example, then some interval estimators and hypothesis tests will be rational belief forming methods, but point estimators will not be.

With this model of belief, let's return to the analysis of scientific knowledge. Unfortunately, scientific knowledge is not true belief acquired via a "good" estimator; this can be shown via a standard Gettier problem. Suppose the medical researcher's new drug is strictly more effective than the conventional treatment, but that, using a "good" estimator, she comes to believe it is strictly less effective. Consider the proposition E, "The efficacy of the new and old drugs differ." If scientific knowledge were true belief acquired via a "good" estimator, then the researcher would know E, but that's unintuitive. Equally problematic, a good estimator for large samples might produce lucky guesses with only one data point, but scientific knowledge rarely arises from a sample of one.

Thus, I say that the researcher's belief in E is *stable* after observing n many patients in a world in which the efficacy equals θ if (1) prior to having collected data, the probability that she would now believe E was high, and (2) the probability that she continues to believe E remains high and approaches one as the data increase. Formally, this means that, if θ represents the actual world and n is the researcher's sample size, then the following condition holds for some number a that is close to zero:

$$\textsc{Stable}_\theta(\hat{e}_n, a) : P_\theta(\hat{e}_k \subseteq E) \geq 1-a \text{ for all } k \geq n, \text{ and } \lim_{k\to\infty} P_\theta(\hat{e}_k \subseteq E) = 1$$

If the researcher's data are misleading, then her beliefs will not be stable, as further evidence will undermine the reasons she currently holds for her beliefs. In *Meno*, Plato argued that knowledge requires stability in a sense, and many philosophers have defended the thesis since.

I now explore two theories of scientific knowledge. The first, which I call CCS, identifies scientific knowledge with true, stable belief that is produced

by a consistent and confident estimator. In other words, let a be some small probability, and say the researcher CCS-*knows*$_a$ a proposition E given her sample x of size n in a world in which the efficacy is θ if the following conditions hold:

> FACTIVITY: $\theta \in E$, i.e., E is true,
>
> BELIEF: $\hat{e}_n(x) \subseteq E$, i.e., the researcher believes E on the basis of her estimator or test,
>
> CONSISTENT$(\hat{e})$ and CONFIDENCE$(\hat{e}_n, a)$, i.e., the researcher's estimator $\hat{e}$ is consistent and confident at sample size n, and finally,
>
> STABLE$_\theta(\hat{e}_n, a)$, i.e., the researcher's belief in E is stable.

The second theory, which I call SP, is intended to be applied only in hypothesis-testing: it identifies knowledge with belief that is produced by a test with low size and high power. Formally, suppose a null hypothesis Θ_0 is tested against a rival Θ_1, and let a and b be probabilities that are sufficiently close to zero (so no more than one half). Say that the researcher SP-*knows*$_{a,b}$ E given her sample x of size n in a world in which the efficacy is θ if FACTIVITY, BELIEF, and the following two conditions hold:

> SIZE$(\hat{e}_n, a)$: $P_\theta(\Theta_0 \not\subseteq \hat{e}_n) \leq a$ for all $\theta \in \Theta_0$, i.e., the researcher's test has low size at sample size n, and
>
> POWER$(\hat{e}_n, b)$: $P_\theta(\Theta_0 \subseteq \hat{e}_n) \leq b$ for all $\theta \in \Theta_1$, i.e., her test is powerful.

SP is very similar to a theory defended by Mayo [1996]. It turns out that SP and CCS closely resemble philosophical theories of knowledge, and so I describe those theories in order to characterize what can be known in scientific contexts.

2.1 Modal Principles for Knowledge

Epistemologists often use one or more of the following three modal conditions to characterize knowledge:[16]

[16] SAFETY is widely defended as a necessary but insufficient condition for knowledge [Pritchard, 2005, Sosa, 1999, Williamson, 2002]. Kvart [2006] defends a probabilistic version of safety (called "KI") similar to one I consider below.Nozick [1981] argues that SENSITIVITY and ADHERENECE are necessary and jointly sufficient for knowledge, and while his theory is now generally rejected [Kripke, 2011], there are probabilistic versions that are

SAFETY: In nearby possible worlds in which S believes that E, the proposition E is true.

SENSITIVITY: In nearby possible worlds in which E is false, S does not believe that E.

ADHERENCE: In nearby worlds in which E is true, S believes E.

I claim that statistical analogs of these three principles are related to the CCS and SP theories of scientific knowledge. To see why, I formalize the principles so that they can be applied to the case involving the medical researcher.

How can one represent possible worlds in statistical settings? An example will motivate my approach. Imagine the medical researcher believes the proposition E, "The drug's efficacy is at least 90%." To determine whether her belief is sensitive, for example, we must assess whether she believes E in worlds in which the efficacy is less than 90%. So the feature of worlds that is necessary for determining whether the researcher's belief is sensitive (or safe, or adherent) is the drug's efficacy.

Thus, I represent possible worlds by real numbers between zero and one. This representation can be understood in at least two ways, and the reader is free to pick her preferred understanding. On one hand, the number θ can represent the possible world that is just like the actual world, except that the drug's efficacy is θ. On the other, θ might represent a set of worlds that are sufficiently similar to the actual world, except that the drug's efficacy in those worlds is θ. Clearly, this representation is adequate only for a *subset* of possible worlds (e.g., worlds in which a drug is being tested). So while I sometimes speak of "all" possible worlds, I really mean "all worlds representable by the statistical model," which in this case, is all numbers between zero and one.

One might object that possible worlds are not always fruitfully represented by numbers because numerical distance is sometimes a poor proxy for distance between possible worlds. I agree. If physicists are estimating the temperature of a substance with discontinuous phase transitions (e.g., from liquid to solid), then it would be inappropriate to represent possible worlds by numerical temperatures. Why? Small temperature differences might correspond to important qualitative distinctions between worlds, and conversely, some larger temperature differences might not track anything

still defended (e.g., Roush [2005]). Black and Murphy [2007] defend a condition similar to Nozick's sensitivity principle, and related principles are employed in contextualist theories of knowledge [DeRose, 1995].

meaningful. Similar remarks apply to estimating the values of fundamental physical constants, such as the electron's charge.

This worry, however, does not undermine the value of my results below for two reasons. First, all of my results hold if possible worlds can be represented by any set Θ, so long as there is a metric d that quantifies the distance between any two worlds in Θ. Thus, it is unnecessary to assume that worlds are numerically representable. Second, many epistemologists already implicitly use numbers to represent sets of possible worlds.[17]

But the objector might continue: are possible worlds plausibly represented by numbers in the running example? If a breast cancer drug's efficacy were different, wouldn't that require human biology to be importantly different? If so, wouldn't small numerical differences fail to track important differences between possible worlds?

Answers to these question will depend on the drug. If possible worlds are representable by single numbers, then biochemistry, animal models, etc. must suggest that the drug's efficacy does not change over time or vary across subpopulations. Otherwise, possible worlds might be better represented by several numbers, which represent the drug's effects at different times or on different subpopulations. Moreover, to motivate the assumptions that possible worlds are representable by real numbers (rather than discrete points) and that numerical distance is an appropriate proxy for distance between worlds, I imagine that the drug's efficacy is a continuous function of *dosage* or of the concentration of some compound in the human body. In contrast, if the nearest possible worlds are ones in which the drug's efficacy differs because of a discrete genetic mutation in the population, then some numbers might not represent nearby possible worlds (e.g., because no plausible mutations changes the efficacy to exactly 92.7%) and numerical distance might be a total inappropriate proxy for distance between possible worlds (e.g., because only remotely plausible mutations change the efficacy a small amount, and plausible mutations change it significantly).

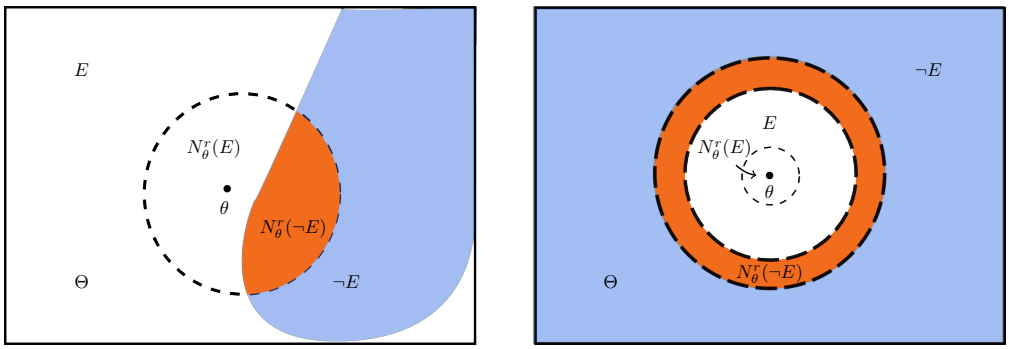
As is standard, I represent propositions by sets of possible worlds. Thus, the proposition represented by E (e.g., the efficacy is at least 93%) is true at a world represented by θ (e.g., a world in which the efficacy is 94%) just in case $\theta \in E$ (e.g., $.94 \in [.93, 1]$).

[17]For instance, Williamson [2002, pp, 18-19,114-115] considers cases in which possible worlds are represented by numerical heights of persons and trees respectively. Hawthorne [2004, p. 45] considers an example involving the number of pounds of salmon a person consumes. Pritchard [2012, p. 179] considers an example in which the numerical distance between a person and bullet is the relevant way of measuring distance between worlds.

SENSITIVITY, ADHERENCE, and SAFETY all concern "near" possible worlds. I remain agnostic about how near is "near enough." All I assume is the following. Take any positive number r and let E be any proposition.[18] Say that a world υ is *E-near-enough$_r$* to the actual world θ if (i) E is true in υ, and (ii) there is no other world η in which E is true such that η is at least r closer to θ. In general, define the set $N^r_\theta(E)$ of "E-near enough$_r$" worlds to θ as follows:

$$N^r_\theta(E) := \{\upsilon \in E : \neg(\exists \eta \in E) d(\theta, \eta) + r \leq d(\theta, \upsilon)\}.$$

An example will illustrate these definitions. Consider the proposition $E = [.9, .95]$ that the drug's efficacy is between 90% and 95%. Suppose the actual efficacy θ is 91%, and let $r = 2\%$ quantify what counts as "near enough." Then $N^r_\theta(E)$ is the set of numbers $[.9, .93)$ between .9 and .93, but excluding .93. Why? If the drug's efficacy is less than .9 or greater than .95, then E is false, and so worlds represented by those numbers cannot count as "nearby" worlds in which E is true. Moreover, if the drug's efficacy is between 93% and 95% inclusive (e.g., if its efficacy is 94%), then although E is true, there is some world η in which E is true and the efficacy is at least $r = 2\%$ closer to the actual efficacy of 91% (e.g., there is a possible world in which the drug's efficacy is $92\% = 94\% - 2\%$). The diagram below illustrates, for two different propositions E, what $N^r_\theta(E)$ and $N^r_\theta(\neg E)$ look like if the set of possible worlds are represented by points in a plane.



If $B_r(\theta) \not\subseteq E$ If $B_r(\theta) \subseteq E$

[18] For the rest of the paper, I use the lower-case Greek letters θ, η, υ, and ζ to represent possible worlds, and I use Θ to denote the set of all possible worlds. I use the lower-case letters like k, m, n to denote natural numbers, whereas r, s, y and z will denote real numbers. I use the letter x to denote data. Like propositions, I ignore the distinction between a world and its representation $\theta \in \Theta$.

If E is a tautology, I will say the E-near-enough$_r$ worlds are nearby *all things considered.* With one modification, one can then restate SENSITIVITY and ADHERENCE using this definition of "near enough."

What modification? In the running example, a possible world is represented by a number, which represents the drug's efficacy. Importantly, the efficacy does not determine which patients survive, but only the probability that a patient survives. Because the researcher's beliefs are determined by which patients survive, therefore, one can determine only what she will *probably* believe in different worlds. To explore SENSITIVITY, ADHERENCE, and SAFETY in statistical settings, therefore, I rephrase the conditions in terms of what the researcher probably believes. For example, SENSITIVITY asserts, "In all near enough worlds in which E is false, it is unlikely that the researcher believes E."[19] How unlikely? Again, the reader may choose any two numbers $a, b > 0$ such that $a, b < 1/2$. I will say a researcher's belief in E is adherent and/or sensitive at sample size n in the world θ if the corresponding condition(s) below holds:

$$\text{ADHERENCE}^r_\theta(E, \hat{e}_n, a)\text{: } P_\upsilon(\hat{e}_n \nsubseteq E) \leq a \text{ for all } \upsilon \in N^r_\theta(E).$$

$$\text{SENSITIVITY}^r_\theta(E, \hat{e}_n, b)\text{: } P_\upsilon(\hat{e}_n \subseteq E) \leq b \text{ for all } \upsilon \in N^r_\theta(\neg E).$$

To state the statistical version of SAFETY precisely, I need one more piece of notation. Informally, a researcher's belief in a proposition E is safe if E is true in all nearby worlds in which it's highly probable that the researcher believes E. How probable is highly probable? Again, the reader may decide. Let $a < 1/2$ be some sufficiently small probability. Then the set of worlds in which it's probable that the researcher believes E after observing n many patients is $\text{BEL}(E, \hat{e}_n, a) = \{\upsilon \in \Theta : P_\upsilon(\hat{e}_n \subseteq E) > 1 - a\}$. Hence, if θ is the actual world, then the "near enough$_r$" worlds in which it's probable that the researcher believes E are represented by the set $N^r_\theta(\text{BEL}(E, \hat{e}_n, a))$, which I will abbreviate $\text{BEL}^r_\theta(E, \hat{e}_n, a)$. To say that E is true in all such worlds is just to say that $\text{BEL}^r_\theta(E, \hat{e}_n, a) \subseteq E$. So one might define SAFETY to be the principle $\text{BEL}^r_\theta(E, \hat{e}_n, a) \subseteq E$.

Unfortunately, this condition is trivially satisfied if there are no nearby worlds in which it's probable that the researcher believes E. After all, the researcher might have luckily arrived at a true belief in the actual world. So in order to ensure that one cannot acquire a safe belief luckily, I stipulate that the researcher's belief in E counts as safe only if it's highly probable in

[19] So my principles are close to what [Kvanvig, 2004] calls the "narrow-scope" (probabilistic) statements of sensitivity, safety, etc.

the actual world that the agent would have come to believe E. Thus, I say the the researcher's belief in E is safe in θ if the following holds:

$$\textsc{safety}^r_\theta(E, \hat{e}_n, c)\text{: } \theta \in \textsc{Bel}^r_\theta(E, \hat{e}_n, a) \subseteq E.$$

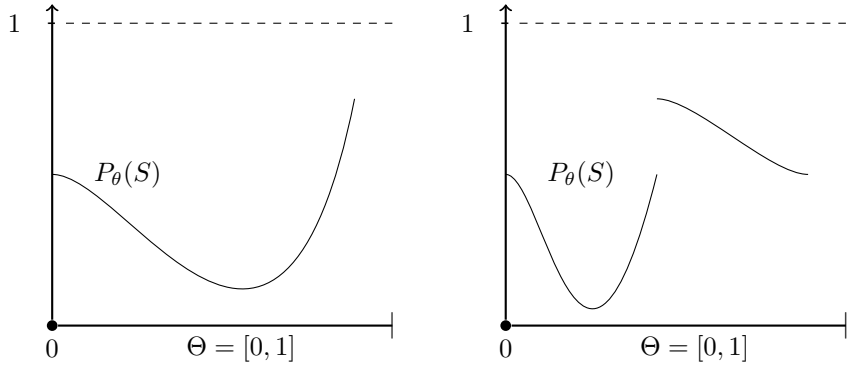
What is the relationship between SENSITIVITY, ADHERENCE, and SAFETY, and the two theories of scientific knowledge above? It turns out that two pairs of theories agree on which propositions are *knowable in principle*. More precisely, let $a > 0$ be some small probability, and say that E is CCS-*knowable*$_a$ in a world θ if there is a data set x and an estimator $\hat{e}$ such that a researcher using the estimator $\hat{e}$ would CCS-know$_a$ (see above) E given data x in θ. Similarly, say a proposition E *can be safely*$_a$ *believed* by the medical researcher in a world θ if there is an estimator $\hat{e}$, a sample size n, and data set x of size n such that (1') the estimator $\hat{e}$ recommends believing E given the data x (i.e., $\hat{e}_n(x) \subseteq E$) and (2') $\textsc{safety}^r_\theta(E, \hat{e}_n, a)$ holds.

Theorem 1 *Suppose there is a consistent estimator with asymptotic confidence greater than* $1 - a$*. Then for all worlds* θ*, a proposition* E *is* CCS-*knowable*$_a$ *in* θ *if and only if* E *can be safely*$_a$ *believed in* θ*.*

See the appendix for a proof. This theorem holds under general conditions, not just in the example involving the medical researcher. Recall that, in the general version of my model, I do not assume that possible worlds are numerically representable, but I do assume that there is a metric d that quantifies the distance between worlds. Moreover, I assume for each world θ, there is some probability distribution P_θ that specifies the chances of obtaining different data sequences in θ. For Theorem 1 to hold in this general model, it is only necessary that roughly, when two worlds are close according to the metric d, then the data one would observe in one world are, with high probability, similar to the data one would observe in the other. I call this assumption *smooth chances*. For example, in a world in which the drug's efficacy is 93%, the researcher's data will, with high probability, be similar to her data in a world in which the efficacy 92.9%. In both types of worlds, about 93% of treated patients will probably survive.

In the example involving the medical researcher, there is an easy visual way of understanding the smooth chances assumption. Let S be any event, e.g., that the researcher observes three of ten patients die or that she believes some proposition E. For each possible world θ, we can plot the probability $P_\theta(S)$ of S occurring. The smooth chances assumption says that, for any event S, the resulting curve will be continuous, in the sense that there are

never any jumps, like the left curve below. All of my results hold in the example about the medical researcher, and they hold in the more general model if smooth chances does.[20]



Smooth Chances | Failure of Smooth Chances

The previous theorem describes the relationship between SAFETY and CCS. The next shows the relationship between Nozick's theory and SP. Say a proposition E is Nozick-knowable$_{a,b}$ if FACTIVITY, BELIEF, ADHERENCE$^r_\theta(E, \hat{e}_n, a)$ and SENSITIVITY$^r_\theta(E, \hat{e}_n, b)$ hold, and say it is SP-knowable if ADHERENCE and SENSITIVITY are replaced by SIZE$(\hat{e}_n, a)$ and POWER$(\hat{e}_n, b)$.

Theorem 2 *Suppose E is tested against its negation $\neg E$ and that some test has (asymptotically) low size and high power. If E is* SP*-knowable$_{a,b}$ in a world θ given data x, then it is Nozick$_{a,b}$ knowable in θ given x.*[21] *Similarly for $\neg E$.*

Theorem 1 is a bit involved, but the proof of Theorem 2 is straightforward. Consider the medical researcher again. Suppose she SP-knows the null hypothesis, so that her belief is based on a test with low size and high power. Because her test has low size, it follows that in *all* worlds (that are

[20] See appendix for a precise statement of smooth chances. Theorems 6 and 8 require the stronger assumption of **uniformly smooth chances**. Although uniformly smooth chances is a strictly stronger assumption, I will not discuss the distinction in depth because (1) in the running example, smooth chances entails uniformly smooth chances because the set of worlds is compact, and (2) almost every applied statistical problem I know has uniformly smooth chances, rendering the distinction relatively unimportant for practice.

[21] The converse is true under mild assumptions. See Proposition 1 in appendix. Also, see Proposition 2, which discusses the relationship between SAFETY and hypothesis tests.

represented in the model) in which the null hypothesis is true, there is a small chance that she will reject it. Hence, the same conclusion holds in all *nearby* worlds in which the null hypothesis is true. So her belief is adherent. Similarly, because her test has high power, in *all* worlds in which the alternative hypothesis is true, there is a small chance she will accept it. As the alternative hypothesis $\neg E$ is true if E is false, it follows that in *all* worlds in which the null hypothesis is false, there is a small chance she will accept it. Hence, in *nearby* worlds in which the null hypothesis is false, there is a small chance she will accept it. So her belief is sensitive. Nozick himself suggested the parallel between his tracking conditions and hypothesis testing:

> *The tracking account of knowledge was formulated without having statistical errors in mind, but conditions 3 [sensitivity] and 4 [adherence] can be (roughly) put as: the person wouldn't commit a type II error about p, and he wouldn't commit a type I error about p.* [Nozick, 1981, p. 260, footnote]

Nozick does not clarify why his tracking conditions are only "roughly" like minimizing Type I and Type II errors. One difference is that statistical power is defined with respect to a particular alternative hypothesis, and the alternative hypothesis need not be the negation of the null. Thus, according to SP, one can only know that the null hypothesis is true *rather than* a particular alternative. That is, SP is a *contrastive* theory of knowledge, whereas Nozick's theory is not.[22]

Importantly, Theorems 1 and 2 do not entail, for example, that CCS and SAFETY agree about what can be known *given a particular sample*. Rather, the theorems concern what can be known in principle, i.e., given *some possible* data set. Thus, a researcher might safely believe E given her current data, but she might not know E according to CCS.[23] Nonetheless, CCS and SAFETY often agree, as do Nozick's theory and SP. Hence, because Nozick's theory violates EC and SAFETY might as well, the next section investigates whether CCS and SP likewise violate EC.

3 Hypothesis Testing and Closure

In this section, I prove that the probabilistic analog of Nozick's theory violates EC when hypotheses are inferred from several independent statistical

[22] For discussions of contrastive theories of knowledge, see [Karjalainen and Morton, 2003] and [Schaffer, 2005]. Thanks to Chris Hitchcock for pointing out that contrastive theories of knowledge, like SP, still might violate closure.

[23] The converse (that CCS entails SAFETY) is true, however. See Theorem 7.

tests; an analogous example works for the SP theory.[24]

Suppose the medical researcher is now interested in testing the efficacy of one hundred different breast cancer treatments. Possible worlds are no longer representable by a single number, but instead, must be represented by a list of one hundred numbers, where each list stands for the efficacy of the hundred treatments. Similarly, propositions are no longer representable by sets of numbers, but by sets of lists of numbers.

For example, consider the proposition $E(n)$: "The n^{th} drug is strictly better than the standard treatment." This proposition is represented by the set of lists in which the n^{th} number is strictly greater than 93%, and the remaining numbers are anything between zero and one (inclusive). Another proposition is $(\exists x)E(x)$: "There is at least one drug that is better than the standard treatment", which is equal to the set of lists in which some number is strictly greater than 93%, and the remaining numbers are anything between zero and one (inclusive).

Suppose the medical researcher conducts one hundred different hypothesis tests - one for each drug. Her null hypothesis in each case is that the n^{th} drug is no more effective than the conventional treatment, i.e., $\neg E(n)$. Suppose that each treatment is tested on two hundred patients, and so the medical researcher has a lot of data. In particular, suppose each statistical test she employs has a size equal to one percent, which recall means that if the n^{th} drug is no more effective than the conventional treatment, then there is at most a one percent chance that she will believe otherwise.

Finally, imagine that the first drug cures breast cancer completely, but that the remaining drugs are equally effective as the conventional treatment (i.e. the actual world is represented by the list in which the first number is one and the remaining numbers are .93). So the propositions $E(1)$ and $(\exists x)E(x)$ are true. I claim that, according to the statistical analog of Nozick's theory of knowledge, the medical researcher knows $E(1)$ and that $E(1) \rightarrow (\exists x)E(x)$, but that she does not know $(\exists x)E(x)$. Therefore, the statistical version of Nozick's theory violates EC. An analogous example works for SP.

Consider $E(1)$ first. The first drug prevents all deaths from breast cancer. Thus, as long as the researcher uses a reasonable statistical test, she will believe the first drug is better than the conventional treatment after treating two hundred patients. To show her belief is adherent, suppose $r = \frac{1}{100}$, and so the nearby worlds in which $E(1)$ is true are ones in which the first drug's efficacy is at least 99%. Because the first drug is very effective in

[24] See Theorem 9 in appendix.

all nearby worlds, she will likewise reject the null hypothesis in all nearby worlds. So her belief is adherent. Note, one may assume that r is larger if the researcher's sample is sufficiently big.

To show her belief is sensitive, consider the size of the test that the researcher employs. In the nearest worlds in which $E(1)$ is false, the first drug is equally effective as the conventional treatment. By assumption, in all such worlds, the researcher will (falsely) reject the null-hypothesis with probability at most one percent. So her belief is sensitive. Thus, according to Nozick's theory of knowledge, the medical researcher knows $E(1)$.

Next I claim the researcher knows $E(1) \rightarrow (\exists x)E(x)$. In fact, it is easy to show that, if the researcher is logically omnicredulous, then she is logically omniscient according to Nozick's theory. This is problematic in general, but it is innocuous in this case. Under any reasonable theory of knowledge, every competent adult knows that "If the first drug is effective, then at least one drug is effective."

Why does the medical researcher fail to know $(\exists x)E(x)$ (according to Nozick's theory)? It suffices to show her belief in $(\exists x)E(x)$ is not sensitive. In the closest world in which $(\exists x)E(x)$ is false, all one hundred treatments are equally effective as the conventional treatment. Call this world θ_0. In θ_0, there is a one percent chance that the medical researcher falsely believes that any given treatment is more effective than the conventional one. Nonetheless, because she conducts one hundred independent tests, and because she expects to get a wrong answer about one in every one hundred times, it is highly probable that the researcher will believe that some drug is effective. In fact, there is only a thirty-six percent chance that the researcher will believe $\neg(\exists x)E(x)$ in θ_0, even though it is true. So in one of the nearby worlds in which $(\exists x)E(x)$ is false, namely θ_0, it is highly probable that the researcher falsely believes $(\exists x)E(x)$. Hence, her belief is not sensitive.[25]

SAFETY likewise violates EC when several independent statistical tests are conducted. To see why, imagine now that the first drug is more effective than the conventional treatment, but only slightly so. Hence, the world in which all one hundred drugs are equally effective as the conventional treatment is nearby, all things considered. I claim that the medical researcher

[25]Surprisingly, according to Nozick's theory, the researcher knows the *conjunction* $(\exists x)E(x)\&E(1)$. Thanks to Arif Ahmed for pointing this out to me. Note that Kripke [2011] and Kvanvig [2004]'s examples show that knowledge, for Nozick, is not closed under conjunction elimination or universal elimination. Nozick [1981, pp. 227-228] gladly admits these rules violate EC. In contrast, the above example concerns existential introduction, which shows Nozick [1981, p. 236] was wrong to claim, "Knowledge, almost always, will be closed under existential generalization."

safely believes $E(1)$ and $E(1) \rightarrow (\exists x)E(x)$, but she fails to safely believe $(\exists x)E(x)$.

Why is the researcher's belief in $E(1)$ safe? Because she employs a test with size 1%, in all nearby worlds in which the first drug is no more effective than the conventional treatment, there is less than a one percent chance that the researcher believes $E(1)$. So in all nearby worlds in which there is a high chance of believing $E(1)$, the drug is effective. Further, the medical researcher's belief in $E(1) \rightarrow (\exists x)E(x)$ is vacuously safe, as there are no worlds in which the proposition is false. Again, although it is unintuitive to say that a belief in a complex tautology is vacuously safe,[26] the assumption is innocuous in this case as $E(1) \rightarrow (\exists x)E(x)$ is sufficiently simple.

Why is the researcher's belief in $(\exists x)E(x)$ not safe? Recall, the first drug is only slightly effective. So in one nearby world, all one hundred drugs are equally effective as the conventional treatment. In that world, it is highly probable that the medical researcher believes $(\exists x)E(x)$ by the same reasoning as above, and so there is a nearby world in which she falsely believes $(\exists x)E(x)$.

Although this example may seem contrived, performing many statistical tests is common in scientific practice. When a prescription drug is tested for safety, medical researchers conduct dozens of statistical tests to determine whether the drug has various side effects. When biologists investigate hereditary diseases, they test whether occurrence of the disease is associated with hundreds of genes. Examples can be multiplied. If scientists draw inferences from statistical tests, their knowledge (according to the theories above) will not be closed under logical consequence.

One might object that when several statistical tests are run simultaneously, many scientists use "correction factors" (e.g., Bonferroni). For example, some scientists use tests with size $\frac{5}{n}\%$, where n is the number of tests one is conducting. Those scientists do not endorse SP, but rather a variant of the theory – call it SP* – where the size and power required for knowledge depends upon the number of statistical tests conducted. One might argue that if correction factors are chosen carefully, then according to SP*, knowledge is closed under logical consequence. Hence, if SP* is the dominant theory endorsed by scientists, then my example fails to show that EC is violated by theories of scientific knowledge.

However, even if it were standard practice to use correction factors when several statistical tests are conducted *simultaneously* or when tests concern

[26] Pritchard [2007] discusses an alternative version of the safety condition that avoids the consequence that all necessary truths are safely believed.

related phenomena, it is not standard practice to correct tests conducted at different times in widely different experimental settings. For example, when a physicist tests whether a new particle has been found, she does not adjust the size of her test because her friend in the psychology department is testing whether implicit biases exist. Intuitively, acquiring knowledge does not require employing correction factors in these cases. So if it is plausible, then SP* does not require always using correction factors.

Importantly, my argument that SP, SAFETY, and Nozick's theory violate EC does not require the assumption that *one* person or one lab *simultaneously* conducts many statistical tests. Nor does it require the assumption that the tests concern related phenomena. The same argument works if one test concerns gas prices, another the Higgs-Boson, a third implicit biases, and so on. In fact, the most questionable assumption, namely that the statistical tests are probabilistically independent, is more plausible in these cases. So my argument shows that SP* – and any theory requiring only "intuitive" uses of correction factors – violates EC. Because the scientific community conducts hundreds of statistical tests daily, EC is routinely violated according to widely-endorsed theories of scientific knowledge, even those that require "correction factors."

The attentive reader will note I have yet to discuss CCS. That is because it satisfies EC. Suppose a researcher knows that E entails F, and assume she knows E according to CCS. I claim she also knows F. Because the researcher believes E and she is logically omnicredulous, she also believes F. Because the researcher knows E, her belief is entailed by the output of a confident and consistent estimator. Because E entails F, it follows that F is likewise entailed by the output of a confident and consistent estimator. Because she is logically omnicredulous and her belief in E is stable, her belief in F is also stable. So the researcher's belief in F satisfies all the conditions of CCS.

It might be surprising that CCS satisfies EC, whereas SAFETY does not. After all, by Theorem 1, every proposition that is knowable in principle according to one theory is knowable in principle according to the other. However, the theories do not always agree on what can be known given a fixed data set. So it is possible that (i) E entails F, (ii) E is safely believed given one's data, (iii) both E and F are CCS-known given one's data, but (iv) F can be safely believed only given *some other data set.*

4 Confidence, Consistency, and Closure

The previous sections present a puzzle. I have argued that theories of scientific knowledge violate EC. However, it would be odd to hear a physicist say, "Recent improvements in technology allow us to estimate the charge of an electron to thirty decimal places, but our evidence does not determine the same value to seven decimal places." Why do violations of EC arise in some areas of science but not others? In this section, I argue that knowledge acquired via estimators with high confidence is closed under entailment, according to all of the theories considered thus far. Violations of EC will be rare, therefore, when confidence intervals are available.

One definition is necessary. In previous sections, I discussed cases in which a proposition follows from a *single* premise. However, a theory of knowledge may satisfy single-premise EC, and yet admit cases in which (i) an agent knows both E_1 and E_2 but (ii) fails to know the conjunction $E_1 \& E_2$.

How should one represent an agent whose beliefs are closed under multi-premise rules of inference? Clearly, a proposition follows from $E_1, \ldots E_n$ just in case it follows from the conjunction $E_1 \& E_2 \& \ldots \& E_n$. Thus, a proposition (considered as a set of worlds) follows from a (possibly infinite) set of propostions $\{E_i\}_{i \in I}$ just in case it is a superset of the intersection $\cap_{i \in I} E_i$. Hence, I say that a theory of knowledge satisfies EC if (i) it satisfies single-premise EC and (ii) if a set of propositions are known, then their conjunction/intersection is also known.

As I argued above, Nozick's theory and SAFETY both violate EC. Interestingly, only single-premise inferences cause difficulties for SAFETY and SENSITIVITY, and only conjunction introduction causes difficulties for ADHERENCE.

Theorem 3 *True sensitive belief is closed under conjunction introduction. Formally, let $\{E_i\}_{i \in I}$ be a (possibly infinite) set of propositions. If $\theta \in E_i$ and $\text{SENSITIVITY}^r_\theta(E_i, \hat{e}_n, b)$ holds for all $i \in I$, then $\text{SENSITIVITY}^r_\theta(\cap_{i \in I} E_i, \hat{e}_n, b)$.*

Theorem 4 *True safe belief is closed under conjunction introduction if the conjunction is probably believed. Formally, let $\{E_i\}_{i \in I}$ be a (possibly infinite) set of propositions. Suppose $\theta \in E_i$ and $\text{SAFETY}^r_\theta(E_i, \hat{e}_n, a)$ holds for all $i \in I$. If $P_\theta(\hat{e}_n \subseteq \cap_{i \in I} E_i) > 1 - a$, then $\text{SAFETY}^r_\theta(\cap_{i \in I} E_i, \hat{e}_n, a)$.*

Theorem 5 *True adherent belief is not closed under conjunction introduction. Formally, there is an experiment, a possible world θ, a sample size n, and a finite set of propositions $\{E_i\}_{i \leq k}$ such that (1) $\theta \in \cap_{i \leq k} E_i$, (2)*

$\text{ADHERENCE}^r_\theta(E_i, \hat{e}_n, a)$ *holds for all* $i \leq k$*, but (3)* $\text{ADHERENCE}^r_\theta(\cap_{i\leq k} E_i, \hat{e}_n, a)$ *does not hold.*

Theorem 6 *Suppose* E *entails* F *and that* E *is known in Nozick's sense in a world* θ *given data* x*. Under one technical assumption,* F *is adherently believed. Formally, suppose* $\theta \in E$*,* $\hat{e}_n(x) \subseteq E$*,* $\text{ADHERENCE}^r_\theta(E, \hat{e}_n, a)$*,* $\text{SENSITIVITY}^r_\theta(E, \hat{e}_n, b)$*, and finally* $E \subseteq F$*. If* $B_r(\theta)$ *is topologically connected, then* $\text{ADHERENCE}^r_\theta(F, \hat{e}_n, a)$*.*

I sketch the argument for Theorem 3; see the appendix for the remaining proofs. Suppose $E_1, E_2, \ldots, E_n$ are true, and assume the researcher's belief in each proposition is sensitive. Now consider the conjunction $F = E_1 \& E_2 \& \ldots \& E_n$. Clearly, the conjunction is true. Further, because I have assumed the researcher is logically omnicredulous, she also believes F. Is her belief sensitive? Yes.

Consider an arbitrary nearby world υ in which F is false. Because F is false in υ, at least one conjunct E_j is false. As υ is a nearby world in which F is false, there are no worlds in which F is false that are significantly closer to the actual world than υ is. But if E_j is false, then the conjunction F is also false. So there are no worlds in which E_j is false that are significantly closer to the actual world than υ is. Thus υ is "near enough" with respect to $\neg E_j$. Because the researcher's belief in E_j is sensitive, she does not believe it in υ. Because she is logically omnicredulous, she does not believe the conjunction F in υ. Because υ was arbitrarily chosen, in every nearby world in which the conjunction F is false, the researcher fails to believe at least one conjunct, and so she does not believe F. So her belief in F is sensitive.

Theorem 6 is perhaps surprising because Kripke [2011] shows that one can know (in Nozick's sense) a conjunction without adherently believing both conjuncts. Kripke imagines a box containing two slits through which a photon might pass. Suppose there is a detector plate behind the right slit but not the left one. Imagine that Mary, a physicist, believes "a photon was emitted" if and only if the detector is activated. Finally, suppose that a photon passes through the right slit activating the detector and causing Mary to believe a photon was emitted. According to Nozick's theory, Mary knows "a photon was emitted, and it passed through the right slit." However, her belief in the proposition "a photon was emitted" is not adherent: in nearby worlds in which the photon passes through the left slit, she does not believe any photon had been emitted.

There is no direct conflict between my theorem and Kripke's example, however. Theorem 6 requires two additional assumptions: smooth chances,

which I have already discussed, and the assumption that the set of nearby worlds, all things considered, is *connected*.

The connectedness assumption says that the set of nearby worlds cannot be divided into two parts L and R such that all of the worlds in L are some fixed positive distance away from all worlds in R. This assumption holds when worlds are representable by sets of real numbers with no "gaps." For example, the set of numbers between zero and one has no gaps, but if we remove all numbers between one eighth and one half, for example then the resulting set has a gap. Connectedness is plausible in scientific applications in which, like the running example, possible worlds are representable by one or more real numbers. But the assumption will fail when possible worlds are best represented by discrete quantities or when nearby possible worlds can be divided into chunks that are qualitatively different (e.g., when some important event happens in some nearby worlds but not others).

There are, therefore, two ways of explaining Kripke's example. First, one might argue that, for any reasonable measure of distance between worlds, the set of nearby worlds is not connected, as the set of worlds L in which the photon passes through the left slit is a fixed distance away from the set of worlds R in which the photon goes right. This is plausible because L worlds are qualitatively different from R worlds: in R worlds, Mary observes something; in L worlds, she doesn't.

But there are ways of quantifying distance under which L worlds and R worlds are arbitrarily close. Suppose that we measure the distance between worlds by how many nanometers apart the trajectories of the photon are. If the two slits can be arbitrarily close, then L worlds and R worlds are arbitrarily close. So the "connectedness" assumption is satisfied. However, the smooth chances assumption is now violated: some L worlds are arbitrarily close to R worlds, and yet the data Mary observes in the two types of worlds are completely different by stipulation. Theorem 6 shows that Kripke's example must be understood in one of these two ways, for otherwise logical consequences of known propositions are adherently believed.[27]

Importantly, the previous four theorems are true of all estimators, i.e., for all belief-formation methods. That is, although safe, sensitive, and adherent belief may not be closed under some rules of inference, it is closed under others, regardless of how the premises of the rules came to be believed.

Are there belief-formation methods that ensure that *all* deductive in-

[27]Kripke's other counterexamples use what I call *second-order* propositions. A second-order proposition asserts something about a first-order proposition E, such as "E and I correctly believe E" or "E and no evidence contradicts E." In my model, I consider only first-order propositions about the drug's efficacy, and I ignore second-order propositions.

ferences preserve knowledge? In some scientific settings, yes. Note that a proposition E is false in a world θ precisely if θ is not an element of E. So if there is a high probability that the actual world is a member of the strongest proposition an agent believes, then there is a high probability that an agent has no false beliefs. And high confidence guarantees there is such a high probability in every world. Thus, if one employs an estimator with high confidence, then there is a low probability that one has *any* false beliefs.[28]

Notice SENSITIVITY requires only that one not believe E in nearby worlds in which it is false. That is, it only requires a low probability of believing a *particular* false proposition. Similarly for SAFETY. So estimators with high confidence always recommend safe and sensitive beliefs. In other words:

Theorem 7 *For any proposition E and any world θ, if* $\textsc{Confident}(\hat{e}_n, a)$, *then* $\textsc{safety}^r_\theta(E, \hat{e}_n, a)$ *and* $\textsc{sensitivity}^r_\theta(E, \hat{e}_n, a)$.

An immediate corollary of Theorem 7 is the following: if the researcher safely (or sensitively) believes E, then she safely (sensitively) believes any proposition F that is entailed by E (as the researcher is logically omnicredulous). Formally, for any two propositions E and F, if $\textsc{Confident}^r_\theta(\hat{e}_n, a)$, $\textsc{safety}^r_\theta(E, \hat{e}_n, a)$, and $E \subseteq F$, then $\textsc{safety}^r_\theta(F, \hat{e}_n, a)$. Similarly for SENSITIVITY. The corollary immediately follows from the theorem because the corollary, like the theorem, is a universally-quantified conditional; the corollary has the same consequent as the theorem, and its antecedent is strictly stronger.

Theorem 7 has at least three important consequences. First, it explains why violations of EC are rare in science. When estimating an unknown quantity, like the charge of an electron, scientists often report confidence intervals. Hence, known propositions (according to a theory that identifies knowledge with safe or sensitive belief) will be closed under entailment.

Second, it explains when deduction creates knowledge. Theories of knowledge that violate EC are often criticized because a mathematical proof or deductive argument from true premises need not extend one's knowledge. Theorem 7 entails that, when one forms beliefs using an estimator with high confidence, then deduction does increase knowledge in Nozick's sense, just as many have claimed.

Third, consider a theory $\textsc{ccs}^+$ that adds SAFETY to the conditions of CCS. By Theorem 1, it follows that every proposition that is in principle $\textsc{ccs}^+$ knowable is safely-believable and vice versa. In the previous section,

[28] Again, the "any" quantifies only over propositions about the quantity being estimated (e.g., the drug's efficacy), and recall one can avoid false beliefs trivially by endorsing only uninformative propositions (e.g., that the drug's efficacy is between zero and 100%).

I argued that CCS satisfies EC. Together with theorem 7, that entails that CCS$^+$ also satisfies EC, as safe belief is closed under entailment if it arises from confident estimators, as CCS$^+$ requires. The moral is that, even if SAFETY violates EC, one can add requirements to safe belief so as to avoid counter-intuitive failures of EC, while at the same time not changing the set of propositions that are, in principle, knowable.

Although high confidence guarantees that sensitive belief is closed under entailment, Nozick's theory requires ADHERENCE also. Recall that adherent belief is not, in general, closed under conjunction introduction. Nonetheless, *uniformly consistent* estimators guarantee that adherent belief is eventually closed under conjunction, just as high confidence guarantees sensitive belief is closed under single-premise inferences.

Consider the running example. Recall, the researcher's estimate is called consistent if it approaches the drug's true efficacy. But the researcher might not know when her guess is close. Uniform consistency, in contrast, allows her to say how much data is required for her guess to be within any margin of error. Formally, $\hat{e}$ is uniformly consistent if $\lim_{n\to\infty} \inf_{\theta\in\Theta} P_\theta(\hat{e}_n \subseteq B_r(\theta)) = 1$ for all $r > 0$. Some standard confidence intervals are uniformly consistent.

Suppose the researcher uses a uniformly consistent estimator, and assume she truly and adherently believes a finite a collection of propositions. The next theorem asserts that it becomes increasingly probable that she believes the conjunction of those propositions, and moreover, that whenever she does believe the conjunction, her belief will be adherent.

Theorem 8 *Let* $\mathbf{F}$ *be a finite set of propositions and* θ *the actual world. If* $\hat{e}$ *is uniformly consistent and* $\text{ADHERENCE}^r_\theta(F, \hat{e}_n, a)$ *for all* $F \in \mathbf{F}$*, then there is* $n_0 \geq n$ *such that* $P_\theta(\hat{e}_k \subseteq \cap\mathbf{F}) > 1 - a$ *and* $\text{ADHERENCE}^r_\theta(\cap\mathbf{F}, \hat{e}_k, a)$ *for all* $k \geq n_0$*.*

Together, theorems 7 and 8 entail that if there is a uniformly-consistent estimator with high asymptotic confidence, then the set of Nozick-knowable propositions is closed under known entailment. The same holds for SAFETY, but the estimator need only be consistent.

5 Conclusion: EC, Obligation, and Permission

When granting permission, delegating obligation, and/or assigning praise and blame, many institutions, I claim, tacitly assume the truth of EC. Whereas anti-skeptical hypotheses play no role in assessments of praise and

blame, scientific knowledge does. Thus, violations of EC in science have additional philosophical importance.

In what ways do institutions tacitly assume EC? Codified duties and permissions are rarely specific. Instead, laws and professional codes of conduct require one to recognize instances of general principles. For example, Rule 1.2 of the American Bar Association's guide for professional conduct requires, "A lawyer shall not counsel a client to engage, or assist a client, *in conduct that the lawyer knows* is criminal or fraudulent" [my emphasis],[29] but the rule does not provide an exhaustive list of circumstances under which that obligation is triggered. Similar remarks apply to codified permissions. When granting power of attorney, an agent cannot provide an exhaustive list of actions that may be taken on her behalf. Instead, one must infer from some particular fact φ that a general fact ψ holds, and then infer they are obliged and/or permitted to act.[30]

In general, here are two intuitively plausible schema about the relationship between knowledge, obligation, and permission. The first says that if an agent (i) knows φ, (ii) knows φ entails ψ, and (iii) is obliged to perform an action a when ψ obtains, then (iv) she is obliged to a. Here, φ is some specific fact (e.g., cigarettes cause cancer), ψ is some general fact (e.g., cigarettes cause harm), and a is an action that is obligatory when the general principle obtains (e.g., place a warning sticker on a label). The second principle is identical, except obligation is replaced by permission.

$$\frac{K(\varphi)\&K(\varphi \to \psi) \quad K(\psi) \to O(a)}{O(a)} \qquad \frac{K(\varphi)\&K(\varphi \to \psi) \quad K(\psi) \to P(a)}{P(a)}$$

These are the type of inference schema that I conjecture are implicitly endorsed in professional and legal settings in which we need to grant permissions and obligations. But neither schema is valid unless EC holds. Moreover, neither can be jettisoned for a schema in which the first premise is replaced by $K(\psi)$, as doing so would require creating exhaustive lists of circumstances under which one is obliged (or permitted) to act. Because (1) implicit standards for scientific knowledge violate EC and (2) many practices presume EC when using scientific knowledge to assign blame and praise, we ought to either revise scientific standards or revise our practices for assigning blame

[29] See [Garwin, 2015].

[30] Some professional codes of conduct also explicitly contain epistemic duties. Modern versions of the the Hippocratic oath, for instance, say, "I will respect the hard-won scientific gains of those physicians in whose steps I walk, and gladly share such knowledge as is mine with those who are to follow" [Lasagna, 2016].

and praise The latter seems less feasible than the former.

Can problematic violations of EC be avoided by requiring scientists to use uniformly estimators with high confidence? Unfortunately, no. In some scientific settings, estimators with high confidence do not exist [Bahadur and Savage, 1956, Donoho, 1988, Tibshirani and Wasserman, 1988], and there are independent reasons to avoid confidence intervals in certain settings [Berger and Wolpert, 1988, Howson and Urbach, 2005]. Thus, we need alternative theories of scientific knowledge that do not violate EC and that are compatible with existing practices for assigning responsibility, permission, and so on. Although I have not proposed such a theory, my hope is that scientifically-minded epistemologists and philosophically-minded statisticians might employ my examples as test cases for future theories of knowledge and that they might use my model and techniques in the development of those theories.

6 Acknowledgements

Thanks to Sam Fletcher, Greg Gandenberger, Kasey Genin, Jonathan Livengood, and Dan Malinsky for helpful suggestions with earlier drafts of this paper. I also benefited greatly from discussions with Colin Marshall, Michael Raven, Ian Schnee, and Justin Zacek. I received helpful feedback from audiences at the 2014 Western Canadian Philosophical Association, the 2015 Formal Epistemology Workshop, and the 2016 meeting of the central division of the American Philosophical Association. Special thanks are due to Sara Aronowitz, Mark Schroeder, and Audrey Yap for providing comments on those presentations, and I am grateful to Arif Ahmed, Adrian Currie, Chris Hitchcock, Hanti Lin, and Chris Stephens for probing questions. Finally, thanks to the editors and two anonymous referees for detailed suggestions about how to improve the content and presentation of my arguments.

References

Marc Alspector-Kelly. Why safety doesnt save closure. *Synthese*, 183(2): 127–142, 2011.

Raghu Raj Bahadur and Leonard J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122, 1956.

James Orvis Berger and Robert L. Wolpert. The likelihood principle. In *Ims Lecture Notes-Monograph*, volume 6. Institute of Mathematical Statistics, 1988.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, New York, second edition, 1986.

Tim Black and Peter Murphy. In defense of sensitivity. *Synthese*, 154(1): 53–71, 2007.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates, second edition, 1988.

Keith DeRose. Solving the skeptical problem. *The philosophical review*, pages 1–52, 1995.

David L. Donoho. One-sided inference about functionals of a density. *The Annals of Statistics*, 16(4):1390–1420, 1988.

F.I. Dretske. Epistemic operators. *The journal of philosophy*, 67(24):1007–1023, 1970.

Jean-Baptist du Prel, Bernd Rhrig, Gerhard Hommel, and Maria Blettner Blettner. Choosing Statistical Tests. *Deutsches Ärzteblatt*, 107(19):343–348, 2010.

Allan Franklin. *Shifting standards: Experiments in particle physics in the twentieth century*. University of Pittsburgh Press, 2014.

Andrew Garwin. Model Rules of Professional Conduct: Table of Contents | The Center for Professional Responsibility, 2015. URL `http://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/model_rules_of_professional_conduct_table_of_contents.html`.

Ronald N. Giere. The epistemological roots of scientific knowledge. *Minnesota Studies in the Philosophy of Science*, 6:212–261, 1975.

Ronald N. Giere. Empirical probability, objective statistical methods, and scientific inquiry. In *Foundations of probability theory, statistical inference, and statistical theories of science*, pages 63–101. Springer, 1976.

Gerd Gigerenzer. *The empire of chance: How probability changed science and everyday life*, volume 12. Cambridge University Press, 1990.

John Hawthorne. *Knowledge and lotteries*. Oxford University Press, 2004.

Wesley H. Holliday. Epistemic Closure and Epistemic Logic I: Relevant Alternatives and Subjunctivism. *Journal of Philosophical Logic*, 44 (1):1–62, February 2015. ISSN 0022-3611, 1573-0433. doi: 10.1007/s10992-013-9306-2.

Colin Howson and Peter Urbach. *Scientific reasoning: the Bayesian approach*. Open Court, Chicago, 2005.

Antti Karjalainen and Adam Morton. Contrastive knowledge. *Philosophical Explorations*, 6(2):74–89, 2003.

Saul A. Kripke. *Philosophical troubles: collected papers*, volume 1. Oxford University Press, 2011.

Jonathan L. Kvanvig. Nozickian Epistemology and the Value of Knowledge. *Philosophical Issues*, 14(1):201–218, October 2004.

Igal Kvart. A probabilistic theory of knowledge. *Philosophy and Phenomenological Research*, 72(1):1–43, 2006.

Louis Lasagna. Hippocratic Oath, Modern version, 2016. URL `http://guides.library.jhu.edu/c.php?g=202502&p=1335759`.

Steven Luper. The Epistemic Closure Principle. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information, fall 2012 edition, 2012. URL `http://plato.stanford.edu/archives/fall2012/entries/closure-epistemic/`.

Deborah G. Mayo. *Error and the growth of experimental knowledge*. University of Chicago Press, 1996.

C. Mayo-Wilson. Estimated True Belief. *Unpublished Manuscript*, 2017.

Peter Murphy. A strategy for assessing closure. *Erkenntnis*, 65(3):365–383, 2006.

Robert Nozick. *Philosophical explanations*. Harvard University Press, 1981.

Duncan Pritchard. *Epistemic luck*. Oxford University Press, 2005.

Duncan Pritchard. Anti-luck epistemology. *Synthese*, 158(3):277–297, 2007.

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

Sherrilyn Roush. *Tracking truth: Knowledge, evidence, and science*. Clarendon Press Oxford, 2005.

Sherrilyn Roush. Closure on Skepticism. *Journal of Philosophy*, 107:243–256, 2010.

Jonathan Schaffer. Contrastive Knowledge. In Tamar Szabo Gendler and John Hawthorne, editors, *Oxford studies in epistemology*, volume 1, pages 235–271. Oxford University Press, 2005.

Frank L. Schmidt. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1(2):115–129, 1996.

Peter Sedlmeier and Gerd Gigerenzer. Do studies of statistical power have an effect on the power of studies? *Psychological bulletin*, 105(2):309–316, 1989.

Ernest Sosa. How to defeat opposition to Moore. *Nous*, 33(13):141–153, 1999.

Robert Tibshirani and Larry A. Wasserman. Sensitive parameters. *Canadian Journal of Statistics*, 16(2):185–192, 1988.

Timothy Williamson. *Knowledge and its limits*. Oxford University Press, 2002.

A Notation

For any $k \in \mathbb{N}$ and any set T, let T^k denote the k-fold Cartesian product of T with itself. Let $T^k_* = \{t \in T^k : t_i \neq t_j \text{ if } i \neq j\}$ be those sequences of length k that have distinct coordinates. For any metric space W, let $\mathbb{B}(W)$ denote the Borel algebra, and let $B_r(w)$ be the open ball of radius r around $w \in W$. Given a sequence of random variables $\langle X_t \rangle_{t \in T}$ and a vector $\vec{t} = \langle t_1, t_2, \ldots, t_k \rangle \in T^k$, we use $X_{\vec{t}}$ to denote the vector $\langle X_{t_1}, \ldots, X_{t_k} \rangle$. If f is a probability distribution and X is a random variable, I write $X \sim f$ to abbreviate the claim that X has distribution f.

B Experiments

An **experiment** $\mathbb{E}$ is a triple $\langle \langle \Theta, d \rangle, T, \langle f^\theta_{\vec{t}} \rangle_{\theta \in \Theta, k \in T, \vec{t} \in T^k_*} \rangle$ where

- $\langle \Theta, d \rangle$ is a metric space,
- Either $T = \mathbb{N}$ or $T = \{1, 2, \ldots, n\}$ for some $n \in \mathbb{N}$, and
- For all $\theta \in \Theta$, all $k \in T$, and all $\vec{t} \in T^k_*$, the function $f^\theta_{\vec{t}} : \mathbb{R}^k \to \mathbb{R}^{\geq 0}$ is a probability distribution over $\mathbb{R}^k$, and moreover, for each $\theta \in \Theta$, the sequence $\langle f^\theta_{\vec{t}} \rangle_{k \in T, \vec{t} \in T^k_*}$ satisfies the consistency conditions of Kolmogorov's extension theorem [Billingsley, 1986, pp. 507-508].

When $\vec{t} = \langle t \rangle$ is a sequence of length one, we write f^θ_t instead of $f^\theta_{\vec{t}}$. From Kolmogorov's extension theorem [Billingsley, 1986, p. 515], it follows that for every experiment, there is a probability space $\langle \Omega, \mathcal{F}, P \rangle$ such that for each $\theta \in \Theta$, there is a sequence $\langle X^\theta_t \rangle_{t \in T}$ of random variables $X^\theta_t : \Omega \to \mathbb{R}$ such that $X^\theta_{\vec{t}} \sim f^\theta_{\vec{t}}$ for all $k \in T$ and all $\vec{t} \in T^k_*$.

I call elements of Θ **worlds** and members of $\mathbb{B}(\Theta)$ **propositions** or **events**. The set Ω represents possible outcomes of an experiment, and the values of $\langle X^\theta_t \rangle_{t \in T}$ represent the **observed data**. T represents the order in which observations are made or recorded. The probability measure P specifies how likely one is to observe particular types of data. Rather than write $P(X^\theta_t \in E)$, I follow the statistical convention of writing $P_\theta(X_t \in E)$.

Example: Assume one is interested in the bias of a coin. Then the set of possible worlds is $\Theta = [0, 1]$, and one measure of distance between worlds is $d(\theta_1, \theta_2) = |\theta_1 - \theta_2|$. To model infinitely many flips of the coin, let $T = \mathbb{N}$. For each $k \in T = \mathbb{N}$ and $\vec{t} \in T^k_*$, let $f^\theta_{\vec{t}} : \mathbb{R}^k \to \mathbb{R}$ be a binomial distribution $Bin(k, \theta)$, i.e., the unique function such that $f^\theta_{\vec{t}}(b) = 0$ if b is not a binary sequence, and otherwise $f^\theta_{\vec{t}}(b) = \theta^{\#(b)}(1-\theta)^{k-\#(b)}$ where $\#(b)$ is the number of ones/heads in the sequence. Let $\mathbb{E}_B$ be the experiment described here. The subscript "B" stands for "Bernoulli."

□

Given an experiment $\mathbb{E}$, define an infinite series of pseudo-metrics $d_n : \Theta^2 \to \mathbb{R}^{\geq 0}$ as follows:

$$d_n(\theta, \upsilon) = \sup_{F \in \mathcal{F}^n} |P_\theta(\langle X_t \rangle_{t \leq n} \in F) - P_\upsilon(\langle X_t \rangle_{t \leq n} \in F)|$$

Say an experiment has **smooth chances** if for all $n \in T$, the identity map from $\langle \Theta, d \rangle$ to $\langle \Theta, d_n \rangle$ is continuous. Say it has **uniformly smooth chances** if each map is uniformly continuous.

C Estimators and Smooth Chances

An **estimator** $\hat{e} = \langle \hat{e}_n \rangle_{n \in T}$ is a sequence of functions $\hat{e}_n : \mathbb{R}^n \to \mathbb{B}(\Theta)$ such that $\{\omega \in \Omega : \hat{e}(\langle X_t^\theta(\omega) \rangle_{t \leq n}) \subseteq E\} \in \mathcal{F}$ for all $E \in \mathbb{B}(\Theta)$, $n \in T$, and $\theta \in \Theta$. Thus $P_\theta(\hat{e}_n \subseteq E) := P(\{\omega \in \Omega : \hat{e}(\langle X_t^\theta(\omega) \rangle_{t \leq n}) \subseteq E\})$ is well-defined for all $\theta \in \Theta$ and $E \in \mathbb{B}(\Theta)$. Because $\Theta \setminus \{\theta\} \in \mathbb{B}(\Theta)$ for all $\theta \in \Theta$, it follows that $P_\theta(\theta \in \hat{e}_n) := P_\theta(\neg\{\hat{e}_n \subseteq \Theta \setminus \{\theta\}\})$ is likewise well-defined for all θ and all n.

Example: Let $\mathbb{E} = \mathbb{E}_B$ be the Bernoulli experiment above. Suppose one is interested in estimating the bias of the coin. Define $\hat{e}(x) = \{\overline{x}\}$ for all $x \in \mathbb{R}^n$, where $\overline{x} = \frac{1}{n} \sum_{1 \leq i \leq n} x_i$ is the fraction of "heads" flips. It is easy to see that $\hat{e}$ satisfies the appropriate measurability conditions, and so what are typically called "point estimators" are estimators in my sense.

Although size and power are typically well-defined only for hypothesis tests, they can applied to other estimators. Let $\Theta_0, \Theta_1 \in \mathbb{B}(\Theta)$ be disjoint. For every estimator $\hat{e}$, define a test $\hat{e}^\tau$ such that

$$\hat{e}^\tau(x) = \text{ the smallest element } E \text{ of } \textsc{Ans} \text{ such that } \hat{e}(x) \subseteq E.$$

Define the size and power of $\hat{e}$ to be that of its corresponding test.

D Statistical Knowledge from Finite Data

For the remainder of the appendix, I assume:
Assumption: a, b and r are positive real numbers, and $a, b < \frac{1}{2}$.

I use $\mathcal{J}$ as a variable for sets containing one or more of the conditions ADHERENCE, CONSISTENT, etc. Let $\mathcal{N} = \{\textsc{adherence}_\theta^r(E, \hat{e}_n, a), \textsc{sensitivity}_\theta^r(E, \hat{e}_n, b)\}$, $\mathcal{S} = \{\textsc{safety}_\theta^r(E, \hat{e}_n, a)\}$, $\textsc{sp} = \{\textsc{Size}(\hat{e}_n, a), \textsc{Power}(\hat{e}_n, b)\}$, and

$$\textsc{ccs} = \{\textsc{Consistent}(\hat{e}), \textsc{Confidence}(\hat{e}_n, a), \textsc{Stable}_\theta(\hat{e}_n, a)\}.$$

Here, $\mathcal{N}$ stands for "Nozick" and $\mathcal{S}$ stands for SAFETY. I write $\theta, \hat{e}, x \models^{\mathbb{E}}_{a,b,r} K_\mathcal{J}(E)$ if $\theta \in E$, $\hat{e}(x) \subseteq E$, and each condition in $\mathcal{J}$ holds. Say a proposition E is *$\mathcal{J}$-knowable at θ* if there are x and $\hat{e}$ such that $\theta, \hat{e}, x \models^{\mathbb{E}}_{a,b,r} K_\mathcal{J}(E)$. Say *$\hat{e}\mathcal{J}$-knowable at θ* if there is x such that $\theta, \hat{e}, x \models^{\mathbb{E}}_{a,b,r} K_\mathcal{J}(E)$. When they are clear from context or don't matter, I drop the superscript $\mathbb{E}$ and the subscript r, and I write $\theta, \hat{e}, x \models_{a,b} K_\mathcal{J}(E)$. Because CCS and $\mathcal{S}$ do not contain the parameter b, I also drop the subscript b when possible.

E Knowable Propositions

By abuse of notation, define $d(E_1, E_2) = \inf\{d(\theta_1, \theta_2) : \theta_1 \in E_1, \theta_2 \in E_2\}$. The first two lemmas are proven in [Mayo-Wilson, 2017].

Lemma 1 *Suppose that E is $\mathcal{N}$-knowable at θ and $\mathbb{E}$ has uniformly smooth chances. Then there is some $s > 0$ such that $B_s(\upsilon) \subseteq E$ for all $\upsilon \in N^r_\theta(E)$. If $B_r(\theta)$ is connected, then $B_r(\theta) \subseteq E$.*

Lemma 2 *Suppose that $\mathbb{E}$ has smooth chances, and let $\hat{e}$ be a consistent estimator with asymptotic confidence $c > a$. The E is $\mathcal{S}$-knowable at θ if and only if θ is an interior point of E.*

Lemma 3 *Suppose that $\mathbb{E}$ has smooth chances. If $\theta, \hat{e}, x \models_a K_{\text{CCS}}(E)$, then θ is an interior point of E.*

Proof: Suppose for the sake of contradiction that θ is **not** an interior point of E and $\theta, \hat{e}, x \models_a K_{\text{CCS}}(E)$. By definition of CCS-knowable, the estimator $\hat{e}$ has confidence at least $1 - a$ at sample size n, i.e., $P_\upsilon(\upsilon \in \hat{e}_n) \geq 1 - a > \frac{1}{2}$ for all $\upsilon \in \Theta$, where n is the length of x. Let $0 < s < \frac{1}{2} - \inf_{\upsilon \in \Theta} P_\upsilon(\upsilon \in \hat{e}_n)$. As $\mathbb{E}$ has smooth chances, there is $q > 0$ such that $d_n(\theta, \upsilon) < s$ if $d(\theta, \upsilon) < q$.

Because θ is not an interior point of E, there is some $\eta \in \neg E$ such that $d(\theta, \eta) < q$. Because $\hat{e}$ has confidence at least $1-a$ at sample size n, we know that $P_\eta(\eta \in \hat{e}_n) \geq 1 - a$. Because $d(\theta, \eta) < q$, it follows that $d_n(\theta, \eta) < s$ and so $P_\theta(\eta \in \hat{e}_n) \geq 1 - a - s > \frac{1}{2}$. This entails that $P_\theta(\hat{e}_n \subseteq E) < \frac{1}{2} < 1 - a$ because $\eta \notin E$. But then E is not stably believed in θ at sample size n, and so E is not CCS knowable at θ at sample size n, contradicting assumption.

Lemma 4 *Suppose that $\mathbb{E}$ has smooth chances, and let $\hat{e}$ be a consistent estimator with asymptotic confidence greater than $1 - a$. Then E is* CCS*-knowable at θ if and only if θ is an interior point of E.*

Proof: *1* entails *2* by Lemma 3. So it only remains to show that *2* entails *1*. Suppose θ is an interior point of E. Thus, $\theta \in E$. The estimator $\hat{e}$ has sufficiently high asymptotic confidence and is consistent by assumption. So it only remains to be shown that there is some sample point x such that E is stably believed from x onward.

Because θ is an interior point of E, there exists some $q > 0$ such that $B_q(\theta) \subseteq E$. As $\hat{e}$ is consistent, one can pick a sample size n_0 large enough so that $P_\theta(\hat{e}_n \subseteq B_q(\theta)) > 1 - a$ for all $n \geq n_0$. Since the event $\{\hat{e}_{n_0} \subseteq E\}$ has positive probability under P_θ, there is a data sequence x of length n_0

such that $\hat{e}_n(x) \subseteq E$. We claim E is stably believed from x onward in θ, i.e., $\textsc{Stable}_\theta(\hat{e}_{n_0}, a)$. Why? Because, $B_q(\theta) \subseteq E$, it follows that $P_\theta(\hat{e}_n \subseteq E) > 1 - a$ for all $n \geq n_0$ by choice of n_0. Moreover, as $\hat{e}$ is consistent:

$$\lim_{n\to\infty} P_\theta(\hat{e}_n \subseteq E) \geq \lim_{n\to\infty} P_\theta(\hat{e}_n \subseteq B_q(\theta)) = 1.$$

Theorem 1 *Suppose that $\mathbb{E}$ has smooth chances, and suppose there is a consistent estimator with asymptotic confidence greater than $1 - a$ for $\mathbb{E}$. Then a proposition E is* CCS*-knowable if and only if it is $\mathcal{S}$-knowable.*

Proof: Follows immediately from lemmas 2 and 4.

Theorem 2 *Let $\Theta_0, \Theta_1 \in \mathbb{B}(\Theta)$ be a partition of Θ, and let $\hat{e}$ be a* ***binary*** *hypothesis test. If $\theta, \hat{e}, x \models_{a,b} K_{\text{SP}}(\Theta_0)$, then $\theta, \hat{e}, x \models_{a,b} K_{\mathcal{N}}(\Theta_0)$.*

Proof: The proof is sketched in the body of the paper; we repeat it here in symbols. Since $\theta, \hat{e}, x \models_{a,b} K_{\text{SP}}(\Theta_0)$, we know $\theta \in \Theta_0$ and $\hat{e}_n(x) \subseteq \Theta_0$. By $\textsc{Size}(\hat{e}_n, a)$, we know $P_\upsilon(\Theta_0 \not\subseteq \hat{e}_n) \leq a$ for all $\upsilon \in \Theta_0$. Because $\hat{e}$ is binary, we know that $\{\hat{e}_n \not\subseteq \Theta_0\} = \{\Theta_0 \not\subseteq \hat{e}_n\}$. Thus $P_\upsilon(\hat{e}_n \not\subseteq \Theta_0) \leq a$ for all $\upsilon \in \Theta_0$. Because $N^r_\theta(\Theta_0) \subseteq \Theta_0$, we obtain that $P_\upsilon(\hat{e}_n \not\subseteq \Theta_0) \leq a$ for all $\upsilon \in N^r_\theta$, i.e., , $\textsc{adherence}^r_\theta(E, \hat{e}_n, a)$ holds.

Because Θ_0 and Θ_1 partition Θ, it follows that $N^r_\theta(\neg\Theta_0) = N^r_\theta(\Theta_1) \subseteq \Theta_1$ and so $\textsc{sensitivity}^r_\theta(E, \hat{e}_n, b)$ follows from $\textsc{Power}(\hat{e}_n, b)$ and the fact that $\hat{e}$ is binary. Thus, $\theta, \hat{e}, x \models_{a,b} K_{\mathcal{N}}(\Theta_0)$.

Proposition 1 *Let Θ_0 and Θ_1 be connected, disjoint subsets of $\mathbb{R}$. Suppose $\hat{e}$ is a binary hypothesis test such that $d(\theta, \Theta_i) \mapsto P(\hat{e}_n \subseteq \Theta_i)$ is decreasing for all n, all $\theta \notin \Theta_i$, and all $i \in \{0, 1\}$. Let $j \in \{0, 1\}$. Suppose there exists some $\theta \in \Theta_j$ such that $\{\upsilon \in \Theta_j : d(\upsilon, \Theta_{1-j}) \leq d(\theta, \Theta_{1-j})\} \subseteq N^r_\theta(\Theta_j)$ and $\theta, \hat{e}, x \models_{a,b} K_{\mathcal{N}}(\Theta_j)$.*

1. *If $j = 0$, then $\theta, \hat{e}, x \models_{a,b} K_{\text{SP}}(\Theta_0)$. In fact, Θ_0 is $\hat{e}\text{SP}_{a,b}$-knowable at υ at sample size n for all $\upsilon \in \Theta_0$.*

2. *If $j = 1$, then $\theta, \hat{e}, x \models_{b,a} K_{\text{SP}}(\Theta_1)$. In fact, Θ_1 is $\hat{e}\text{SP}_{b,a}$-knowable at υ at sample size n for all $\upsilon \in \Theta_1$.*

Proof: For *1*, because $\theta, \hat{e}, x \models_{a,b} K_{\mathcal{N}}(\Theta_0)$, it follows that $\theta \in \Theta_0$ and $\hat{e}_n(x) \subseteq \Theta_0$, where n is the length of x. So it suffices to show the size of $\hat{e}_n$ is no more than a and its power is at least $1 - b$. Consider size first. Pick $\zeta \in \Theta_0$. We want to show $P_\zeta(\Theta_0 \not\subseteq \hat{e}_n) \leq a$. Because $\hat{e}$ is a binary hypothesis test, it suffices to show that $P_\zeta(\hat{e}_n \not\subseteq \Theta_0) \leq a$.

There are two cases to consider: either $d(\zeta, \Theta_1) \leq d(\theta, \Theta_1)$ or not. In the former case, by assumption, $\zeta \in N^r_\theta(\Theta_j)$. By ADHERENCE, it follows that $P_\zeta(\Theta_0 \nsubseteq \hat{e}_n) \leq a$. In the latter case, our assumption entails $P_\zeta(\hat{e}_n = \Theta_1) \leq P_\theta(\hat{e}_n = \Theta_1)$. Because $\hat{e}$ is a binary hypothesis, it follows that $P_\zeta(\hat{e}_n \nsubseteq \Theta_0) \leq P_\theta(\hat{e}_n \nsubseteq \Theta_0)$. Now $P_\theta(\hat{e}_n \nsubseteq \Theta_0) \leq a$ by ADHERENCE, as $\theta \in N^r_\theta(\Theta_0)$. Thus, $P_\zeta(\hat{e}_n \nsubseteq \Theta_0) \leq a$. Since we've considered both cases, we've shown $P_\zeta(\hat{e}_n \nsubseteq \Theta_0) \leq a$ for all $\zeta \in \Theta_0$, as desired.

For power, we first note that because Θ_0 and Θ_1 are connected, disjoint subsets of $\mathbb{R}$, it follows that either (i) $\theta_0 < \theta_1$ for all $\theta \in \Theta_0$ and $\theta_1 \in \Theta_1$, or (ii) $\theta_1 < \theta_0$ for all $\theta \in \Theta_0$ and $\theta_1 \in \Theta_1$.

Pick $\zeta \in \Theta_1$. We must show that $P_\zeta(\Theta_0 \subseteq \hat{e}_n) \leq b$. Because $\hat{e}$ is binary, it suffices to show that $P_\zeta(\hat{e}_n \subseteq \Theta_0) \leq b$. There are two cases to consider: either $\zeta \in N^r_\theta(\Theta_1)$ or not. In the former case, $P_\zeta(\hat{e}_n \subseteq \Theta_0) \leq b$ by SENSITIVITY. In the latter case, it follows from either (i) or (ii) that $d(\zeta, \Theta_0) > d(\upsilon, \Theta_0)$ for all $\upsilon \in N^r_\theta(\Theta_1)$. So by assumption, $P(\hat{\zeta}_n \subseteq \Theta_0) \leq P_\upsilon(\hat{e}_n \subseteq \Theta_0)$ for all $\upsilon \in N^r_\theta(\Theta_1)$. Finally, by SENSITIVITY, $P_\upsilon(\hat{e}_n \subseteq \Theta_0) \leq b$ for all $\upsilon \in N^r_\theta(\Theta_1)$. Thus, $P_\zeta(\Theta_0 \subseteq \hat{e}_n) \leq b$, as desired.

To prove the second statement in *1*, let $\upsilon \in \Theta_0$. To show that Θ_0 is is $\hat{e}\text{SP}_{a,b}$-knowable at υ at sample size n, we first note that $\upsilon \in \Theta_0$ by assumption and that we've already shown that $\hat{e}$ has the appropriate size and power at sample size n. So it suffices to find some data sequence y such that $\hat{e}_n(y) \subseteq \Theta_0$. Let $y = x$. This completes the proof of *1*. Statement *2* is proven similarly.

Proposition 2 *Let $\Theta_0, \Theta_1 \in \mathbb{B}(\Theta)$ be a partition of Θ, $r > 0$, and $\hat{e}$ be a **binary** hypothesis test. Then:*

1. *Assume $\hat{e}$ has power at least a. For all worlds $\theta \in \Theta_0$ and all data sequences x, if $\hat{e}(x) \subseteq \Theta_0$, then $\theta, \hat{e}, x \models_a K_\mathcal{S}(\Theta_0)$.*

2. *Assume $\hat{e}$ has size no greater than $1 - a$. For all worlds $\theta \in \Theta_1$ and all data sequences x, if $\hat{e}(x) \subseteq \Theta_1$, then $\theta, \hat{e}, x \models_a K_\mathcal{S}(\Theta_1)$.*

Proof: For *1*, it suffices to show that $\text{SAFETY}^r_\theta(E, \hat{e}_n, a)$ holds. Let n be the length of x, and let $\upsilon \in \text{BEL}^r_\theta(\hat{e}_n, \Theta_0, a)$. It must be shown that $\upsilon \in \Theta_0$. Because $\upsilon \in \text{BEL}^r_\theta(\hat{e}_n, \Theta_0, a)$, by definition we have $P_\upsilon(\hat{e}_n \subseteq \Theta_0) > 1 - a > a$. Because $\hat{e}$ has power $\geq a$, we know that $\upsilon \notin \Theta_1$. Because $\{\Theta_0, \Theta_1\}$ is a partition of Θ, it follows that $\upsilon \in \Theta_0$. Statement 2 is proven similarly.

F Epistemic Closure

Say that $\mathcal{J}$ **satisfies single-premise** EC if for all experiments $\mathbb{E}$, all events $E, F \in \mathbb{B}(\Theta)$, all estimators $\hat{e}$, all sample sizes $n \in T$, and all data sequences x of length n:

- $\theta, \hat{e}, x \models^{\mathbb{E}}_{a,b} K_{\mathcal{J}}(E)$ and $\theta, \hat{e}, x \models^{\mathbb{E}}_{a,b} K_{\mathcal{J}}(\neg E \cup F)$ together entail that
- $\theta, \hat{e}, x \models^{\mathbb{E}}_{a,b} K_{\mathcal{J}}(F)$.

Say $\mathcal{J}$ is **closed under (finite) conjunction** if for all experiments $\mathbb{E}$, all (finite) collection of propositions $\{E_i\}_{i \in I}$ such that $\cap_{i \in I} E_i$ is a proposition:

$$\text{If } \theta, \hat{e}, x \models^{\mathbb{E}}_{a,b} K_{\mathcal{J}}(E_i) \text{ for all } i \in I, \text{ then } \theta, \hat{e}, x \models^{\mathbb{E}}_{a,b} K_{\mathcal{J}}(\cap_{i \in I} E_i).$$

Say $\mathcal{J}$ **violates epistemic closure** if either (i) it is not closed under conjunction or (ii) it violates single-premise EC.

Theorem 3 *Let $\{E_i\}_{i \in I}$ be a set of propositions. If $\theta \in E_i$ and* $\text{SENSITIVITY}^r_\theta(E_i, \hat{e}_n, b)$ *holds for all $i \in I$, then* $\text{SENSITIVITY}^r_\theta(\cap_{i \in I} E_i, \hat{e}_n, b)$.

Proof: The proof is sketched in the text. In symbols, the result follows immediately from the fact that:

$$N^r_\theta(\neg(\cap_{i \in I} E_i)) = N^r_\theta(\cup_{i \in I} \neg E_i) \subseteq \cup_{i \in I} N^r_\theta(\neg E_i)$$

and that $\text{SENSITIVITY}^r_\theta(E_i, \hat{e}_n, b)$ holds for all $i \in I$.

Theorem 4 *Let $\{E_i\}_{i \in I}$ be a set of propositions. Suppose $\theta \in E_i$ and* $\text{SAFETY}^r_\theta(E_i, \hat{e}_n, a)$ *holds for all $i \in I$. If $P_\theta(\hat{e}_n \subseteq \cap_{i \in I} E_i) > 1 - a$, then* $\text{SAFETY}^r_\theta(\cap_{i \in I} E_i, \hat{e}_n, a)$.

Proof: By assumption, $P_\theta(\hat{e}_n \subseteq \cap_{i \in I} E_i) > 1 - a$, and so $\theta \in \text{BEL}^r_\theta(\cap_{i \in I} E_i, \hat{e}, a)$. I claim that $\text{BEL}^r_\theta(\cap_{i \in I} E_i, \hat{e}, a) \subseteq B_r(\theta)$. Why ? If $\upsilon \notin B_r(\theta)$, then $d(\theta, \theta) + r = r \leq d(\theta, \upsilon)$. Moreover, $P_\theta(\hat{e}_n \subseteq \cap_{i \in I} E_i) > 1 - a$ by assumption, and so $\upsilon \notin \text{BEL}^r_\theta(\cap_{i \in I} E_i, \hat{e}, a)$ by definition.

Let $\eta \in \text{BEL}^r_\theta(\cap_{i \in I} E_i, \hat{e}, a)$. We must show that $\eta \in \cap_{i \in I} E_i$. Suppose for the sake of contradiction that $\eta \notin \cap_{i \in I} E_i$. Then there exists $j \in I$ such that $\eta \notin E_j$. Because E_j is safely known at θ, it follows that $\eta \notin \text{BEL}^r_\theta(\hat{e}, E_j, a)$. Because $\eta \in \text{BEL}^r_\theta(\cap_{i \in I} E_i, \hat{e}, a)$, it follows that $P_\eta(\hat{e}_n \subseteq E_j) > 1 - a$. So if $\eta \notin \text{BEL}^r_\theta(\hat{e}, E_j, a)$, then there must be some υ such that $d(\theta, \upsilon) + r \leq d(\theta, \eta)$ and $P_\upsilon(\hat{e}_n \subseteq E_j) > 1 - a$. But if $r \leq d(\theta, \upsilon) + r \leq d(\theta, \eta)$, then $\eta \notin B_r(\theta)$, contradicting our assumption that $\eta \in \text{BEL}^r_\theta(\cap_{i \in I} E_i, \hat{e}, a)$ and the observation that $\text{BEL}^r_\theta(\cap_{i \in I} E_i, \hat{e}, a) \subseteq B_r(\theta)$.

Theorem 5 $\mathcal{J} = \{\text{ADHERENCE}\}$ *is not closed under finite conjunction.*

Proof: As in the body of the paper, we consider an experiment in which several hypotheses are tested simultaneously. However, we use a discrete parameter space for simplicity. Let $\Theta = \{0,1\}^{100}$ and let $d(\theta, \eta)$ be the discrete metric. Notice $\mathbb{B}(\Theta)$ is the the power set of Θ. Let $T = \{1, 2, \ldots, 100\}$. For all $t \in T$ and $\theta \in \Theta = \{0,1\}^{100}$, define $f_t^\theta : \mathbb{R} \to \mathbb{R}^{\geq 0}$ as follows.

$$f_t^\theta(x) = \begin{cases} .99 & \text{if } x = \theta_t \\ .01 & \text{if } x = 1 - \theta_t \\ 0 & \text{otherwise, i.e., if } x \notin \{0,1\}. \end{cases}$$

For any $k \in T$ and any $\vec{t} \in T_*^k$, define $f_{\vec{t}}^\theta : \mathbb{R}^k \to \mathbb{R}^{\geq 0}$ as below:

$$f_{\vec{t}}^\theta(x) = \Pi_{j \leq k} f_{t_j}^\theta(x_j).$$

Define $\mathbb{E}$ to be the experiment $\langle \langle \Theta, d \rangle, T, \langle f_{\vec{t}}^\theta \rangle_{\theta \in \Theta, k \in T, \vec{t} \in T_*^k} \rangle$. Informally, we can think of each possible world as a collection of 100 heavily weighted coins. In every world, each of the 100 coins is weighted so it lands on heads with probability .99, or on tails with probability .99. In the experiment $\mathbb{E}$, we flip each coin once. So X_k^θ represents the flip of the k^{th} coin in world θ.

Consider the world $\theta = \langle 1, 1, \ldots, 1 \rangle$, which represents the world in which all coins land heads with probability .99. For each $k \leq 100$, let $E_k = \{\theta \in \Theta : \theta_k = 1\}$, which represents the proposition that the k^{th} coin is biased towards heads. So $\theta \in E_k$ for all k. Let $r > 0$.

For every binary sequence x of length $n \leq 100$, define x^* to be the result of appending $100 - n$ ones to the end of x. For each $n \in T$, define $\hat{e}_n : \{0,1\}^n \to \mathbb{B}(\Theta)$ by $\hat{e}(x) = \{x^*\}$. By definition of $\hat{e}_{100}$, we have

$$P_\upsilon(\hat{e}_{100} \subseteq E_k) = P_\upsilon(X_k = 1) = .99 \text{ for all } \upsilon \in E_k.$$

Hence, if $a = .01$, then $\text{ADHERENCE}_\theta^r(E_k, \hat{e}_n, a)$ for all k. However:

$$P_\theta(\hat{e}_{100} \subseteq \cap_{k \leq 100} E_k) = P_\theta(\cap_{k \leq 100} X_k = 1) = .99^{100} \approx .36.$$

As $\theta \in N_\theta^r(\cap_k E_k)$, it follows that $\text{ADHERENCE}_\theta^r(\cap_{k \leq 100} E_k, \hat{e}_{100}, a)$ fails.

Theorem 6 *Suppose* $\mathbb{E}$ *has uniformly smooth chances. Suppose* $\theta, \hat{e}, x \models_{a,b} K_\mathcal{N}(E)$ *and that* $B_r(\theta)$ *is connected. If* $E \subseteq F$, *then* $\text{ADHERENCE}_\theta^r(F, \hat{e}_n, a)$.

Proof: First, $\theta \in E$ because $\theta, \hat{e}, x \models K_\mathcal{J}(E)$. Because $E \subseteq F$, it follows that $\theta \in F$. Similar reasoning shows $\hat{e}(x) \subseteq F$.

Because $B_r(\theta)$ is connected and E is $\mathcal{N}$-knowable at θ, it follows from Lemma 1 that $B_r(\theta) \subseteq E$. So $N^r_\theta(E) = B_r(\theta)$. Because $E \subseteq F$, it follows that $B_r(\theta) \subseteq F$. Thus, $N^r_\theta(F) = B_r(\theta) = N^r_\theta(E)$.

ADHERENCE$^r_\theta(F, \hat{e}_n, a)$ holds because

$$\begin{array}{lll} \inf\limits_{\upsilon \in N^r_\theta(F)} P_\upsilon(\hat{e}_n \subseteq F) & \geq & \inf\limits_{\upsilon \in N^r_\theta(F)} P_\upsilon(\hat{e}_n \subseteq E) \\ & \text{as} & E \subseteq F \\ & = & \inf\limits_{\upsilon \in N^r_\theta(E)} P_\upsilon(\hat{e}_n \subseteq E) \\ & \text{as} & N^r_\theta(F) = N^r_\theta(E) \\ & > & 1 - a \\ & \text{as} & \text{ADHERENCE}^r_\theta(E, \hat{e}_n, a). \end{array}$$

Theorem 7 *Let $\hat{e}$ be any estimator with confidence at least $1 - a$. Then* SAFETY$^r_\theta(E, \hat{e}_n, a)$ *and* SENSITIVITY$^r_\theta(E, \hat{e}_n, a)$.

Proof: For SENSITIVITY, suppose $\zeta \in \neg E$. Then $P_\zeta(\hat{e}_n \subseteq E) \leq P_\zeta(\zeta \notin \hat{e}_n) < a$ as $\hat{e}$ has confidence $1 - a$. Hence:

$$\sup_{\zeta \in N^r_\theta(\neg E)} P_\zeta(\hat{e}_n \subseteq E) \leq \sup_{\zeta \in \neg E} P_\zeta(\hat{e}_n \subseteq E) < a.$$

For SAFETY, again let $\zeta \in \neg E$. As $\hat{e}$ has confidence $1 - a > a$, it follows that $P_\zeta(\zeta \in \hat{e}_n) > a$. Because $\zeta \notin E$, it follows that $P_\zeta(\hat{e}_n \subseteq E) < 1 - a$. So $\zeta \notin \text{BEL}^r_\theta(E, \hat{e}, a)$.

Corollary 1 *Suppose $\mathbb{E}$ has uniformly smooth chances, that $B_r(\theta)$ is connected and that $\hat{e}$ has confidence $1 - b$ at sample size n. If $E \subseteq F$, x is a data sequence of length n, and $\theta, \hat{e}, x \models_{a,b} K_\mathcal{N}(E)$, then $\theta, \hat{e}, x \models_{a,b} K_\mathcal{N}(F)$.*

Proof: Follows immediately from theorems 6 and 7.

Theorem 8 *Suppose $\mathbb{E}$ has uniformly smooth chances and that $\hat{e}$ is uniformly consistent. Let $\mathbf{F}$ be a finite set of propositions. If* ADHERENCE$^r_\theta(F, \hat{e}_n, a)$ *for all $F \in \mathbf{F}$, then there is $n_0 \geq n$ such that for all $k \geq n_0$ (i) $P_\theta(\hat{e}_k \subseteq \cap\mathbf{F}) > 1 - a$, and (ii)* ADHERENCE$^r_\theta(\cap\mathbf{F}, \hat{e}_k, a)$

Proof: Because $\mathbb{E}$ has uniformly smooth chances, Lemma 1 entails that for each $F \in \mathbf{F}$, there is some positive $s_F < r$ such that $B_{s_F}(\upsilon) \subseteq F$ for all $\upsilon \in N^r_\theta(F)$. Let $s = \min\{s_F : F \in \mathbf{F}\}$, which is positive because $\mathbf{F}$ is finite. So $B_s(\upsilon) \subseteq \bigcap \mathbf{F}$ for all $\upsilon \in N^r_\theta(F)$ and all $F \in \mathbf{F}$.

As $\theta \in \bigcap \mathbf{F}$, it follows that $N^r_\theta(\bigcap \mathbf{F}) = B_r(\theta) \cap \bigcap \mathbf{F} \subseteq B_r(\theta) \cap F = N^r_\theta(F)$ for all $F \in \mathbf{F}$. Hence, $B_s(\upsilon) \subseteq \bigcap \mathbf{F}$ for all $\upsilon \in N^r_\theta(\bigcap \mathbf{F})$. By the uniform consistency of $\hat{e}$, one can pick n_0 large enough such that $P_\upsilon(\hat{e}_k \subseteq B_s(\upsilon)) > 1 - a$ for all $k \geq n_0$ and all words $\upsilon \in \Theta$. Since $B_s(\upsilon) \subseteq \bigcap \mathbf{F}$ for all $\upsilon \in N^r_\theta(\bigcap \mathbf{F})$, it follows that $\sup_{\zeta \in N^r_\theta(\bigcap \mathbf{F})} P_\zeta(\hat{e}_k \nsubseteq \bigcap \mathbf{F}) \leq a$. So $\text{ADHERENCE}^r_\theta(\cap \mathbf{F}, \hat{e}_k, a)$ for all $k \geq n_0$.

Proposition 3 *If $\mathcal{J}$ is equal to $\mathcal{N}$, $\mathcal{S}$, or $\{$SENSITIVITY$\}$, then $\mathcal{J}$ is not closed under single premise entailment. Thus, if $\mathcal{J}$ is a non-empty subset of either $\mathcal{N}$ or of $\mathcal{S}$, then $\mathcal{J}$ violates epistemic closure.*

Proof: See examples in body of the paper.

Theorem 9 SP *violates epistemic closure.*

Proof: Define the experiment $\mathbb{E}$ and estimator $\hat{e}$ as in the proof of Theorem 5. Consider the hypotheses $\Theta_0 = \{0\} \times \{0,1\}^{99}$ and $\Theta_1 = \{1\} \times \{0,1\}^{99}$. By definition of $\hat{e}_{100}$, we have $P_\theta(\hat{e}_{100} \subseteq \Theta_i) = P_\theta(X_0 = i) = .99$ for all $\theta \in \Theta_i$ and all $i \in \{0,1\}$. Hence, the test $\hat{e}^\tau_{100}$ associated with $\hat{e}_{100}$ (See Appendix Section C) has size .01 and power .99 when testing Θ_0 versus Θ_1. Thus, $\hat{e}_{100}$ has size at most .01 and power .99 when testing Θ_0 versus Θ_1. In contrast, let $\Theta^*_0 = \{\theta \in \Theta : \theta_n = 0 \text{ for some } n\}$ and $\Theta^*_1 = \{\theta \in \Theta : \theta_n = 1 \text{ for all } n\}$. Let $\theta = \langle 1, 1, \ldots, 1\rangle$ be the constant vector. Then $P_\theta(\hat{e}_{100} \subseteq \Theta_1) \approx .36.$, and so $\hat{e}_{100}$ has power at most $.36 < 1 - b$ when testing Θ^*_0 vs. Θ^*_1.

Let $\theta = x$ be any sequence that begins with a zero, so that $\theta \in \Theta_0 \subseteq \Theta^*_0$ and $\hat{e}(x) \subseteq \Theta_0$. Then $\theta, \hat{e}, x \models_{a,b} K_{\text{SP}}(\Theta_0)$, but $\theta, \hat{e}, x \not\models_{a,b} K_{\text{SP}}(\Theta^*_0)$.