Practical Epistemology: Essays On What To Think and What To Do

by

Haley Schilling

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2022

Author .....
Department of Linguistics and Philosophy
August 29, 2022

Certified by .....
Roger White
Professor of Philosophy
Thesis Supervisor

Accepted by .....
Bradford Skow
Chair of the Committee on Graduate Students

Practical Epistemology: Essays On What To Think and What To Do

by

Haley Schilling

Submitted to the Department of Linguistics and Philosophy
on August 29, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Philosophy

Abstract

Sometimes, we need evidence in order to act. A jury needs "proof beyond a reasonable doubt" in order to convict a defendant of a crime. A teacher needs to read a student's essay in order to assign a grade. A babysitter needs to know that the sandwich does not contain peanuts, in order to give it to a child with a peanut allergy. The FDA needs "substantial evidence" of the efficacy of a new drug in order to approve it. This dissertation explores the relationship between ethics and epistemology, evidence and practical deliberation, and what to think and what to do.

Chapter 1 develops an account of "proof beyond a reasonable doubt," a standard that is vexingly difficult to pin down. Legal proof is knowledge on the basis of trace evidence that the defendant is guilty. This epistemic norm generalizes to all of our responses and reactive attitudes — and is a challenge to orthodox knowledge norms.

Chapter 2 considers a central issue in the ethics of AI — algorithms are often opaque. This essay characterizes a class of applications for which there is a special moral demand for transparency: algorithms that give people what they deserve, on the basis of what they have done. Explainability is important to assure that the algorithms follow the requisite epistemic norms.

Chapter 3 considers the pragmatic encroachment thesis, the claim that whether S knows p depends on the practical, as well as the epistemic, features of her deliberative context. The essay argues that the ordinary knowledge ascriptions that often motivate the thesis can just as easily undermine thesis, and then develops a contextualist knowledge norm that can account for the data.

Chapter 4 explains how to set significance levels, based on practical considerations. Scientists should set significance levels based on the value of the posterior credences that would result from updating on different results of significance tests.

Thesis Supervisor: Roger White
Title: Professor of Philosophy

Acknowledgements

MIT has been a wonderful philosophical home. So many of you have contributed to this dissertation in seminars, the lounge, and at WIP, MATTI, ERG, and Dissertation Seminar.

Thank you.

For Q&A, feedback on drafts, lively discussion, and the MIT epistemology hivemind, thank you David Balcarras, Allison Balin, Nathaniel Baron-Schmitt, David Builes, Alex Byrne, Yonathan Fiat, Kelly Gaus, Félix-Antoine Gelineau, Cosmo Grant, Jessica Heine, Michele Impagnatiello, Justin Khoo, Abe Mathew, Alex Meehan, Josh Pearson, Anni Räty, Ryan Ravanpak, Agustín Rayo, Miriam Schoenfield, Bob Stalnaker, Eliot Watkins, Mallory Webber, Tyler Brooke-Wilson, Eliza Wells, Xinhe Wu, and Steve Yablo.

Kevin Dorst, thanks for being my buddy, and for excellent comments on drafts. Thomas Byrne, thanks for puzzling through all of this with me, and for reading many drafts. Roger White, Sally Haslanger, Jack Spencer, Kieran Setiya, Caspar Hare, thank you for philosophical wisdom, guidance, encouragement, and open doors.

To Mom, Dad, B, and S, love and gratitude beyond words.

To my father, Tom Schilling, for his love of wisdom

Contents

0 | Introduction

Sometimes, we need evidence in order to act. A jury needs "proof beyond a reasonable doubt" in order to convict a defendant of a crime. A professor needs to read a student's essay in order to assign a grade. A babysitter needs to know that the sandwich does not contain peanuts, in order to pack it in the lunchbox of a child with a peanut allergy. The FDA needs "substantial evidence of efficacy" and "proof of safety" of a new drug in order to approve it.

This dissertation is a collection of essays in practical epistemology — on the relationship between ethics and epistemology, knowledge and deliberation, and what to think and what to do.

Chapter 1, "The Epistemology of Legal Proof," develops an account of the criminal legal standard of "proof beyond a reasonable doubt," which is the evidence that jurors need in order to deliver a guilty verdict. This is an epistemic standard that is vexing to judges and legal epistemologists, both familiar yet notoriously difficult to pin down.

In this essay, I'll defend this thesis: legal proof is knowledge on the basis of trace evidence (i.e., evidence that is causally downstream from the crime) that a defendant is guilty. The jury needs to know that the defendant is guilty, and they also must have a "smoking gun." The fact that knowledge is not sufficient for legal proof is illustrated cases like the following:

> A traffic camera has been recording cars driving through a school zone for the past year. It has caught Annie speeding every weekday for the past year, except for one. On June 17, 2022, the camera batteries were dead, and they did not record the traf-

> fic. Suppose the traffic police review the tape. They are in a position to know from the tape that she has a habit of speeding, but they should not issue her a ticket for speeding on June 17. They can complain: "we know she did it, but unfortunately, we didn't catch her."

This is a surprising result. Knowledge is often taken to be the strongest epistemic state. Many epistemologists hold that if S knows that p, then it is rational for S to act as if p.

Cases multiply outside of the courtroom. A professor must read their student's term paper before assigning a course grade, even if they have had the student in class before and know that they would produce an excellent A+ term paper. An official at a high school swim meet must declare a winner based on what they see in the pool, even they knew in advance that the clear frontrunner would win the race. You may know that your roommate will forget to water the plants while you are away— she is so scatterbrained and always forgets these things — but you can't blame her until you get back home and see that the plants are wilting. When we respond to the actions of others, it is important that we respond *causally* to what they have done, and it follows that our evidence must be caused by what they have done.

Chapter 2, "The Ethics of XAI" considers a central issue in the ethics of artificial intelligence. Algorithmic decision making is ubiquitous — algorithms grade essays, grant parole, evaluate teachers, diagnose cancer, and so on. The algorithms are often opaque. The people whose lives the algorithms affect — and in many cases, the engineers themselves — have no idea how they work. The opacity is a source of significant moral concern, and there is a significant endeavour by computer scientists and engineers to figure out how to create explainable artificial intelligence (XAI)

This essay argues that XAI is important for reasons that extend beyond the bias or accuracy of the algorithms. Specifically, XAI is important to ensure that algorithms are acting for the right reasons. In a broad range of applications, the algorithms are used to give people what

they deserve, on the basis of what they have done. They *respond* to what other people have done — they allocate punishments, rewards, and evaluations. These algorithms must act for the right reasons and with the right evidence. XAI is important to ensure that they do. This class of algorithms is broader may initially meet the eye. One problem with the COMPAS algorithm for parole decisions is that it makes a *prediction* about the future conduct of a parole candidate, where it should be a *reaction* to the candidate's conduct in prison. The problem with COMPAS is not just that it is a "Black Box" it's that it's a "*crystal ball*." It should be a response to what people do, the kind of response for which XAI is especially important.

The essay will then argue that these responses are subject to epistemic norms. In Chapter 1, we saw that our responses are subject to epistemic norms. A similar epistemic norm applies in the case of algorithmic responses. The algorithms must know that the basis for their response holds, and the responses must be caused by what other people have done. XAI is important to ensure that his epistemic norm holds.

Along the way, this essay will connect the dots between opacity and other moral concerns that arise in the ethics of AI literature: the use of proxy variables (for race, gender, and so on), the right to be treated as an individual, the right to appeal decisions, and the threat of algocracy.

Chapter 3, "A Puzzle for Pragmatic Encroachment" considers the *pragmatic encroachment* thesis. According to thesis, whether S knows that p depends on more than the *epistemic* features of her deliberative context (i.e., her evidence, belief forming process, the truth of p, and so on.) It also depends on the *non-epistemic* features of her deliberative context, like her options and their practical stakes.

The classic example: Sarah the babysitter is packing a lunch for a young child, Algernon. The sandwich on the left is peanut butter, the sandwich in the middle is tuna, and the sandwich on the right is almond butter. She knows the sandwich in the middle is tuna. Her roommate told her that the one on the left is peanut butter and the one on the right is almond butter, but she can't tell those sandwiches apart by sight, taste, or smell.

Suppose Algernon likes peanut butter, dislikes almond butter, and tolerates tuna. Which sandwich should Sarah pack? The one on the left. Why? Because she knows the one on the left is peanut butter and the one on the right is almond butter. Now suppose Algernon has a severe peanut allergy. Which sandwich should Sarah pack? The tuna sandwich. Why? Because she doesn't know the one on the left is peanut butter and the one on the right is almond butter. What accounts for the difference between the cases? According to the pragmatic encroachment thesis, the practical stakes make a difference to what she knows.

In this essay, I argue that this case actually creates a puzzle for the thesis. Suppose Sarah gives Algernon the sandwich on the left, the peanut butter butter sandwich, because she hates him and wants him dead. We can blame her: "you knew that was a peanut butter sandwich!" The stakes are the same as in the previous case, but the knowledge ascriptions are different. The essay develops a contextualist account of the pragmatic encroachment thesis that can account for all of this data, one that takes seriously the role of knowledge ascriptions in practical deliberation, and in our ordinary practices of praise and blame.

Chapter 4, "How To Set Significance Levels" considers a parallel thesis to the pragmatic encroachment thesis in the context of science: statistical significance levels should be set based on the practical costs of errors. According to the Standard Account, if a Type I error is costly, scientists should lower significance levels. If a Type II error is costly, scientists should raise the significance levels. In this essay, I develop an account of how to set significance levels. Significance levels should on e that depends on practical considerations, but that's different from the Standard Account.

In this essay, I argue that the Bayesian paradigm provides a better *theory* of scientific rationality than the frequentist paradigm — yet it does not follow that we should "abandon statistical significance" testing or "ban the p-value." Frequentist statistics can still be useful in *practice*. Bayesian researchers face cognitive and spatiotemporal limitations, and they need statistics in order to effectively condense, communicate, and learn from data. This is an important role for

statistics, and one that significance tests play well. It's a practice in the *context of discovery* and not the *context of justification*. To run a statistical test is to condense the data into one-bit of information, or to learn the answer to one "yes/no" question about the data. The value of a question depends on the value of the answers, and so we arrive at the following account of the connection between significance levels and non-epistemic values:

1. How we should set significance levels depends on the value of the posterior credence functions that could result from conditionalization on the significance test.

2. The value of a posterior credence function depends on epistemic and non-epistemic considerations.

I'll then account for how researchers should set statistical significance levels in two cases. First, the classic example: how the FDA should set significance levels as they evaluate drugs for "substantial evidence of efficacy" and "proof of safety." Second, the application for which Neyman-Pearson significance tests were developed: the eugenics project of searching for differences in heritable traits among ethnic groups, and in contemporary "neurosexist" and "scientifically racist" research projects. In the first application, the Standard Account gets things right. In the second application, the Standard Account errs, in ways that may have grave consequences.

1 The Epistemology of Legal Proof

Introduction

A jury needs "proof beyond a reasonable doubt" in order to convict a defendant of a crime. The standard is used in most adversarial legal systems, is enshrined in the 14th amendment of the US constitution, and is widely held to be necessary to secure a just criminal conviction. Yet the standard is notoriously difficult to pin down: it's been characterized as "fundamental and universally familiar ... but in practice it is vexingly difficult to interpret and apply [Whitman, 2008]." The US Supreme Court does not require that any explication of the standard be given to juries in jury instructions, and some jurisdictions prohibit judges from elaborating on or clarifying the standard [Tanford, 1990]. From *People v. Johnson:* "It is difficult to find a plainer or more explicit definition of reasonable doubt than the words themselves, and efforts to do so usually result merely in an elaboration of language without any corresponding amplification of the idea."

Legal epistemologists, including Moss [2018, 2021], Blome-Tillmann [2017], Littlejohn [2020], and Pardo [2010, 2011], have circled in on an account of the standard: legal proof is knowledge. A jury has proof beyond a reasonable doubt just in case they know that the defendant is guilty. In this essay, I'll argue that knowledge is necessary — but not sufficient — for legal proof. The courtroom evidence can support knowledge that the defendant is guilty, without constituting legal proof. What's missing in these cases is *trace evidence*: evidence that is caused by or causally explained by the crime. Legal proof is knowledge on the basis of trace evidence that the defendant

is guilty.

The essay will then defend a general epistemic thesis, for which legal proof provides a vivid illustration. All of our responses to what other people do — punishment, evaluations, praise, blame, and so on — require knowledge on the basis of trace evidence. This is a surprising thesis, one that undermines the orthodox view about the value of knowledge and its role in practical reason: if S knows that p, then S is in an epistemic position to act as if p.

1.1 Knowledge and Legal Proof

In a criminal trial, the prosecutors hold the burden of proof. The defendant has a presumption of innocence: they are considered innocent until proven guilty. The evidence required in order to convict a defendant is substantial, which is due to the serious miscarriage of justice that is a wrongful conviction.[1] From Blackstone, "the law holds that it is better that ten guilty escape than that one innocent suffer" and the opinion in *Winship*: "I view the requirement of proof beyond a reasonable doubt in a criminal case as bottomed on a fundamental value determination of our society that it is far worse to convict an innocent man than to let a guilty man go free." To guard against false convictions, legal proof must entail that the defendant is guilty with a high probability, which provides a desideratum on any account of legal proof, and suggests that knowledge is required for legal proof.[2]

Although legal proof must entail that there is a high probability that a defendant is guilty, legal proof is not merely evidence that supports a high probability that the defendant is guilty. The evidence can support a high probability that a defendant is guilty, without supporting a

[1] See Schwikkard[1999], Laudan [2005], and Tadros [2007] for an in-depth discussion and defense of the presumption of innocence and the prosecutor's burden of proof.

[2] A qualification. The jury must know on the basis of *admissible* evidence that the defendant is guilty. Some evidence is excluded from the courtroom on the grounds that it is prejudicial to the jury (e.g., some character evidence, or evidence of prior convictions). Some evidence is excluded as a matter of policy (e.g., confessions obtained in an interrogation when a defendant was not read their Miranda rights, or from an illegal search). If this evidence is introduced to the jury, the judge will request that the jury disregard it in their deliberations, and they should deliberate as if they were never presented with this evidence.

conviction. This is illustrated in cases of "merely statistical evidence." Here's an example from [Nesson, 1979]:

> **Prisoners:** 24 out of 25 prisoners who are incarcerated on a prison cellblock form a riot and kill a guard. One of the prisoners does not join the riot and runs away and hides. The riot is captured on tape by a security camera, but when prosecutors review the tape, they cannot discern the identities of the prisoners. They are all wearing identical jumpsuits, and the tape is too grainy to identify them by their features. The prosecutors select a defendant at random and charge him with murder.

The evidence in this case supports a high probability that the defendant is guilty.[3] 24 out of 25 of the prisoners participated in the riot, and there is no reason to think any of them in particular were involved in the riot. A conviction in this case may strike you as unjust. If so, you share this intuition with me, judges in actual cases, legal epistemologists, and the general public (psychologists call the unwillingness to convict in this case the "Wells effect.")[4] Evidence can support a high probability in the defendant's guilt without justifying a guilty verdict.

Why doesn't the evidence suffice for a conviction? The defendant on the stand was selected at random, and so the probability he is guilty is .96. This is a high probability — and evidence that supports a .96 probability in the defendant's guilt usually supports a conviction. Suppose that instead of a video tape, a credible (but not infallible) eyewitness testified that the defendant on the stand committed the crime. This testimony may support a .96 probability in the defendant's guilt, but count as legal proof.

The thesis that knowledge is legal proof can explain why the jury does not have sufficient evidence to convict in the *Prisoners* case. The example is similar to so-called "lottery propositions" in the epistemology literature, named for the following example. Suppose you purchase a lottery

[3] I'm using "probability" here, but any graded doxastic state will do, for example, credence or epistemic probability.

[4] See Gardiner [2019] for a summary of the literature in support of this intuition, and Niedermeier [1999] for the psychological literature on the Wells effect.

ticket. Your evidence supports a high probability that you will lose the lottery, but you do not know that your ticket will lose. You should not rip up your ticket or assert "my ticket will lose." Your lack of knowledge is not due to a low probability that the ticket will lose — there is a high probability that you will lose. Instead, it's due to the probabilistic nature of your evidence. If you read the winning numbers in the newspaper, you could know that your ticket lost, even if it does not change the probability much (it was extremely probable that you would lose before you read the paper and extremely probable after.) In the *Prisoners* case, the proposition that the defendant is guilty is similar to the proposition that your lottery ticket will lose. Just as you cannot know that your lottery ticket will lose, the jury cannot know that the defendant is guilty. Knowledge as legal proof can explain why the jury cannot convict, despite the high probability that the defendant is guilty.

Knowledge as legal proof is also an application of the knowledge norm of action — an account of the role that knowledge plays in practical deliberation outside of the courtroom. This is suggested by the jury instructions that are read by judges in some federal and state courts (*Holt v. US*):

> "A reasonable doubt is a doubt based upon reason and common sense — the kind of doubt that would make a reasonable person hesitate to act. Proof beyond a reasonable doubt must, therefore, be proof of such a convincing character that a reasonable person would not hesitate to rely and act upon it in the most important of his own affairs."

The jury instructions equate proof beyond a reasonable doubt with the epistemic state that makes it reasonable to rely on or act upon a proposition in action or assertion. According to many epistemologists, this epistemic state is knowledge. Here's the knowledge norm of assertion, which is defended by Unger [1975], Williamson [2000], and Hawthorne [2004]:

> **Knowledge Norm of Assertion:** S is in an epistemic position to assert p just in case

S knows that p.

Strong arguments can be given for this epistemic norm. First, the necessary condition: If S is in an epistemic position to assert p, then S knows that p. This norm is supported by ordinary judgements about appropriate assertion. If somebody asks S for directions, and she doesn't know which way to go, she should not assert: "head North." In the lottery case, you do not know that your lottery ticket will lose, so you should not assert: "my ticket will lose."

The knowledge norm can also explain what's wrong with quasi-Moorean sentences, like: "it's raining outside, but I don't know that it's raining outside." According to the norm, if S is in a epistemic position to assert that it's raining, then she knows that it is raining, and she should not say that she does not know it's raining.

Next, the sufficient condition: If S knows that p, then S is in an epistemic position to assert p. Here, the caveat that knowledge puts S is in an *epistemic* position to assert p is crucial: S should not assert everything she knows. She probably doesn't have time for that, but also, she would risk saying something rude, or harmful, or so on.

In support of this sufficient condition, note that knowledge is a strong epistemic state. If S knows that p, then: S believes p, S's evidence that supports a high probability that p, p is true, and so on. So if any of these epistemic conditions are sufficient for S to assert p, then knowledge is also sufficient. And to deny this sufficient condition, we would have to accept some strange consequences, for example: "there are things people know but ought not to assert because their epistemic position is not strong enough with respect to those things." Hawthorne [2004] calls this sentence "disturbing."

According to many epistemologists, the knowledge norm of assertion also generalizes to action:

> **Knowledge Norm of Action:** S is in an epistemic position to act as if p just in case S knows that p.

Knowledge as legal proof is an application of the knowledge norms to the juror's guilty verdict. A juror is in an epistemic position to act as if the defendant is guilty (i.e., the juror has legal proof and can deliver a guilty verdict) just in case the juror knows that p.

Finally, there is a meta-argument for knowledge as legal proof. Knowledge and legal proof are both familiar, yet vexing, perplexing, and resistant to explication. Similar necessary or sufficient conditions have been given for both legal proof and knowledge: evidence that supports a high probability in the defendant's guilt [Papineau, 2021], safety [Pritchard, 2018], sensitivity [Enoch et al., 2012] , the elimination of relevant alternatives [Gardiner, 2018], and so on, which are clues that they are one and the same.

1.2 Trace Evidence and Legal Proof

The thesis that knowledge is the standard for legal proof, despite the compelling arguments in its favor, is false. While knowledge may be necessary for legal proof, it is certainly not sufficient. This can be illustrated with the following cases.

> **Traffic:** A traffic camera has been recording cars driving through a school zone for the past year. It has caught Annie speeding every weekday for the past year, except for one. On June 17, 2022, the camera batteries were dead, and they did not record the traffic.

Suppose the traffic police review the tape. They are in a position to know from watching the tape that she has a habit of speeding through the stop signs. The police can infer that she sped through the traffic light on June 17, but they should not issue her a ticket for June 17. They may complain: "we know she did it, but unfortunately, we didn't catch her."

> **Trading:** Bob and Carol are executives at a major energy company, and they are business partners and close confidants. Bob has access to insider information that suggests that the company's stock prices will plummet. Bob tells Carol that he plans to

> sell his stocks, and that it's a perfect crime: his investment transactions are concealed in a maze of off-shore accounts that would be impossible for federal prosecutors to trace. He always follows through on such plans. His phone has been wiretapped by the prosecutors, and they have this confession caught on tape.

Suppose the prosecutors listen to the tape. They are in a position to know that Bob sold his stocks, but they should not press charges. Again, they may complain: "we know he did it, but unfortunately, he left no trace."

The evidence against the suspects in *Traffic* and *Trading* does not constitute legal proof. This may strike you as correct as a matter of moral principles, but it's also borne out in the law. The insufficiency of the evidence in these cases follows from a patchwork of common law principles and the US Federal Rules of Evidence. The Federal Rules of Evidence 404(b) concern "other crimes, wrongs, or acts" and state: "evidence of any other crime, wrong, or act is not admissible to prove a person's character, in order to show that on a particular occasion the person acted in accordance with the character." The evidence of other crimes can be admitted for a narrow range of specific uses: "this evidence may be admissible for another purposes, such as proving motive, opportunity, intent, preparation, plan, knowledge, identity, absence of mistake, or lack of accident." In the *Traffic* case, Annie's speeding on the days other than June 17 does not fall under one of these narrow exceptions, so the evidence would not even be allowed into the courtroom, never mind suffice for a conviction.

The *corpus delicti* (Latin: "body of the crime") rule from common law, codified in almost all US jurisdictions, guards against convictions for crimes that never occurred. From *Wigmore*, in order to convict a defendant, one must prove: "three component parts, first, the occurrence of the specific injury or loss (as, in homicide, a person deceased; in arson, a house burned; in larceny, property missing); secondly, somebody's criminality as the source of the loss — these two together involving the commission of a crime by somebody; and thirdly, the accused identity as the doer of the crime." The third component, the identity of the perpetrator, is the one that is at issue in

cases of merely statistical evidence, where it is clear that a crime has in fact occurred. As a result, it has received much attention in legal epistemology. The first on this list is the *corpus delicti*, which is the component that is missing in the *Traffic* and *Trading* cases. In the *Traffic* case, the prosecutors must show that the crime actually happened, and courts are unequivocal: Annie's speeding habit is not enough. From *United States vs. Woods*: "the exclusive use of prior acts, without more, cannot establish the *corpus delicti*."

Like the evidence of other criminal acts, a defendant's confession alone cannot establish the *corpus delicti*. Common law, following ancient Roman law, holds that in order to establish *corpus delicti*, confessions must be corroborated with evidence *aliunde* (Latin: "from elsewhere"), which is independent, substantial, and corroborating evidence of the confession. In the *Trading* case, the prosecutor's only have Bob's uncorroborated confession, which alone is not enough to establish his guilt. So the evidence in the *Traffic* and *Trading* cases would not suffice for a conviction in actual common law courts, and if a prosecutor presented a judge with this evidence when pressing charges, the judge would be required to dismiss the charges on the grounds of lack of evidence.

The evidence in the *Traffic* and *Trading* cases support knowledge that the defendant is guilty. In the *Traffic* case, we can know that Annie sped through the traffic light on June 17, 2022, because we know she has a habit of speeding. In general, we can know what people have done on the basis of their habits. For example, you can know that your spouse had a cup of coffee this morning, because they have a cup of coffee every morning, even if you left the house this morning before they woke up. You can know that they wore a seatbelt when they drove to work, because they always wear a seatbelt when they drive, even if you were not in the car to witness it. In the *Trading* case, we can know that Bob sold his stocks, because he said he would. In general, we can know what people have done on the basis of what they said they will do. For example, you can that your colleague has sold her bitcoin, because she told you that she was going to at the watercooler. And you can know that she called her mom to wish her a happy birthday, because she said she was going to do that. To deny knowledge in these cases opens the door to skepticism

about the lives of other people.

The courtroom, of course, is no ordinary context. Jury deliberations can be a matter of life and death, and a wrongful conviction is a serious miscarriage of justice. According to the *pragmatic encroachment* thesis, the stakes of the jury's deliberative context can make a difference to what they are in a position to know. Because of the high stakes, more evidence (i.e., a higher credence) is required in order to know that Annie sped through a traffic light on June 17 than to know that your spouse had a cup of coffee this morning. This thesis is compelling in the context of jury deliberations. It is reflected in the jury instructions which state that the evidence for the guilty verdict must be convincing enough that the reasonable person would act on it "in the most important of his own affairs."

Even if we grant the pragmatic encroachment thesis (see Chapter 3), the jury can still know that the defendant is guilty in the *Traffic* and *Trading* cases. Two arguments. First, the evidence is just as probable as the evidence that does support a conviction. A police officer's testimony that Annie was speeding (in lieu of a pattern) could count as legal proof. The police officer's testimony and the pattern of speeding support a similar credence in Annie's guilt. So if the former case counts as legal proof, so does the latter. Second, in the *Traffic* case, the linguistic data suggests that the evidence supports knowledge that the defendant is guilty. The police say, truthfully: "we know she did it, but unfortunately, we didn't catch her." The cases are genuine counterexamples to knowledge as legal proof.

The inference patterns in these examples may seem familiar. They are also used to construct counterexamples to the causal theory of knowledge [Goldman, 1967]. This theory of knowledge was introduced as a response to the Gettier problem: suppose you see Brown driving a Ford, you you conclude that Brown owns a Ford, and then infer the disjunction that Brown owns a Ford or Brown is in Barcelona. Unbeknownst to you, Brown does not drive a Ford (you saw him driving a rental), and in a remarkable coincidence, Brown just so happens to be in Barcelona. You have a justified true belief that Brown owns a Ford or Brown is in Barcelona, but you do not know this.

The causal theory of knowledge gives the following explanation: your belief that Brown owns a Ford or Brown is in Barcelona is not caused by the state of affairs that makes that belief true, namely, that Brown is in Barcelona.

This theory of knowledge, despite any initial plausibility, is false. It admits of counterexamples: it rules out knowledge on the basis of some inference patterns. The knowledge that an unobserved emerald is green can result from enumerative induction: you observe many emeralds, and then can conclude that all emeralds are green. But your belief that some unobserved emerald is green is not caused by the fact that the unobserved emerald is green. Future predictions, based on past events, can also be a source of knowledge. We can know that the sun will rise tomorrow or that weather tomorrow will be sunny. But these beliefs are not caused by tomorrow's sunrise or sunshine. The counterexamples to knowledge as legal proof are also counterexamples to the causal theory of knowledge.

This suggests that what's missing in the *Traffic* and *Trading* cases is evidence that stands in the right causal relationship with the defendant's crime. What's missing is what's referred to in legal scholarship as "trace evidence," which is evidence that is caused by, or causally explained by, a crime. It's the proverbial "smoking gun," which leads us to the following account of legal proof:

> **Legal Proof:** The criminal legal standard of "proof beyond a reasonable doubt" is knowledge on the basis of trace evidence that the defendant is guilty.

A few clarifications are in order. First, the requirement that legal proof be knowledge that is based on trace evidence does not require that the knowledge be based entirely on trace evidence. A jury may be presented with evidence about events that are not caused by the crime (or that even prompted the crime), for example, to establish the defendant's motive. This evidence may be crucial for the prosecution's case and should be presented to a jury and be part of the basis on which a juror arrives at a guilty verdict.

Second, the trace evidence must be *evidence that the jury has* for the defendant's guilt. The trace evidence must be presented to the jury. If the jury knows that the defendant is guilty, they will also know that there's some trace evidence out there that would confirm the defendant's guilt. In *Trading*, the jury could reason: we know Bob sold his stocks, so there is some paper trail of the transaction in the records of the offshore banks, even though we do not have access to them. So, the trace evidence must be evidence that's presented to the jury. Similarly, the trace evidence must be *evidence* for the crime. If the prosecutors have a very large stack of papers with all of the stock trades from Bob's company on the day he dumped his stocks, this is not evidence that Bob committed the crime, and it is not part of the basis for which the jury comes to know that Bob committed the crime.

The phrase "*corpus delicti*" (body of the crime) may suggest that trace evidence is a certain kind of tangible physical evidence, the kind that could be photographed or shown to the jury (Exhibit A: photographs of the body, Exhibit B: the bloody knife). This is part of the popular understanding of the law (it's sometimes called the "no body, no crime" rule), but this is a misunderstanding. Omissions, for example, can be trace evidence of crimes. A disappearance is trace evidence of a murder, even if the body is never recovered.

This is an *externalist* norm. Whether or not a juror knows that a defendant is guilty is not determined entirely by their internal mental states. Jurors in different cases could be physical duplicates, while one knows that the defendant is guilty, and thereby have legal proof, while the other does not. Knowledge is factive: in order for a jury to know that a defendant is guilty, it must be true that the defendant is guilty, which is not internal to the defendant. Similarly, the demand for trace evidence is also externalist. Evidence may appear to a jury to be trace evidence, but actually not be. Suppose a traffic cop testifies that he saw Annie speed through the traffic light on June 17, but he actually got the dates mixed up and wasn't actually on the scene that day to witness her speeding. To the jury, the evidence appears to be credible trace evidence, and they may think that they based a guilty verdict on trace evidence, but they in fact did not. If this

happens, the defendant has a complaint: the crime played no causal role in the verdict, but the jury is not to blame. They have an excuse: they reasonably believed that they had trace evidence for the crime.

So, we can give separate conditions that account for when a juror is blameworthy for failing to follow the norm. Here are conditions on blameworthiness given in terms of reasonable belief, although similar conditions could be given in different epistemic terms (e.g., in terms of justification or excuses). A juror is blameworthy for delivering a guilty verdict if they do not reasonably believe that they have knowledge with trace evidence that the defendant is guilty. A juror is blameworthy for delivering a not-guilty verdict if they reasonably believe that they have knowledge on the basis of trace evidence that the defendant is guilty.

The claim here is that legal proof is knowledge on the basis of trace evidence that a defendant is guilty, and just trace evidence alone is not sufficient for a just conviction. Sorensen [2006] defends a "causal theory of verdicts," a riff on Goldman's causal theory of knowledge, using cases similar to *Traffic* and *Trading*. He argues that knowledge is dispensable in the courtroom setting: the correct causal structure between a crime and a guilty verdict is all that justice requires. I disagree. First, there is still the problem of "merely statistical evidence," which the requirement of trace evidence does not solve. An appeal to causation in cases of merely statistical evidence is tempting to some, and I think my account of legal proof explains why: eyewitness testimony is a canonical example of trace evidence. But the requirement of trace evidence does not rule out a conviction on the basis of merely statistical evidence. In the *Prisoners* case, there is trace evidence of the crime, specifically, the videotape of the crime being committed. The causal theory of verdicts bites the bullet: a conviction based on merely statistical evidence alone suffices for a conviction.

Second, there are Gettier cases, which are cases that Sorenson takes to support the causal theory of verdicts, but I think undermine it. Here's a case, similar to one that he gives:

Framed: George kills his brother. He leaves the dead body and a bloody knife with

his DNA on it at the crime scene. A police officer investigating the murder hates George and sees this as an opportunity to frame him for murder, unaware of the fact that George was the actual culprit. So he wipes the knife clean and plants George's DNA on it.

A juror with this evidence has a justified true belief in the evidence, but if they found George guilty, there would be an injustice. Why? According to the causal theory of verdicts, it's because there has been a breakdown in the "usual" causal chain from crimes to verdicts, and this undermines the justice of a verdict.

This case draws out an important difference between the causal theory of verdicts and my account of legal proof. I think the jurors do have trace evidence in this case: they have a dead body. So, the causal theory of verdicts requires a causal connection between a crime and a conviction that is much stronger than knowledge on the basis of trace evidence.

I do not think this verdict is unjust because the jurors lack evidence that is causally downstream from the crime. I think the problem is that they do not know that George is the culprit. (Also, George is a victim of another injustice. the police officer wronged him by fabricating evidence, but I'm going to set this aside as a distracting feature of the case). In the following case, the jurors have the exact same trace evidence for a crime, but know that the defendant is guilty, and this allows them to reach a guilty verdict:

> **Framed (again):** Harry kills his brother. He leaves the dead body and a bloody knife with his DNA on it at the crime scene. A police officer investigating the murder hates Harry and sees this as an opportunity to frame him for murder, unaware of the fact that Harry was the actual culprit. So he wipes the knife clean and plants Harry's DNA on it.

Like George, Harry has a similar complaint against the police officer for trying to frame him for a crime. Unlike George, the jury should convict Harry of a crime. This conviction would be just.

The only difference between the two cases? In the first case, the jury does not know that the defendant is guilty, but in the second case, they do. They know that Harry is the only person who had access to the crime scene.

Just like in George's case, in Harry's case, there is at least some evidence that stands in an unusual causal relationship with the crime: the bloody knife. But this doesn't undermine the justice of the verdict. In many trials, there will be information presented to the jury that appears to be trace evidence, caused by the crime in the usual way, that's actually not. A witness for the prosecution may have a false memory, and it may appear that they are giving credible eyewitness testimony, even if they are not. But this doesn't undermine the justice of a guilty verdict, as long as the jury is still able to know on the basis of trace evidence that the defendant is guilty. This is supported by actual legal practice: a convicted defendant will only be granted a new trial upon the discovery of new evidence if it is "likely to result in a different verdict" or if it "casts real doubt on the justice of the verdict," and not if it turns out to contradict any of the evidence presented to the jury.

1.3 Other Epistemic Conditions?

The lack of a causal connection between crime and conviction is my explanation for the deficiency in the evidence in the *Traffic* and *Trading* cases, but this section will consider other necessary (and in some cases, sufficient) conditions on legal proof that have been defended by legal epistemologists, and will argue that they are unable to account for the data. What about modal epistemic conditions, like safety or sensitivity? These conditions have been raised as necessary conditions for legal proof, often to explain why merely statistical evidence is not sufficient for a conviction.[5]

Here is the safety condition:See Pardo [2018] for an argument for the safety condition for legal proof

[5]For defenses of the sensitivity condition as the explanation for the insufficiency of merely statistical evidence, see [Enoch et. al, 2012]

Safe Verdicts: In the the nearby worlds in which a defendant is convicted, the defendant is guilty.

An informal gloss, or alternative formulation: the jury could not have easily been wrong about the defendant's guilt. This is a kind of "anti-luck" condition that could rule out knowledge in Gettier cases. In our earlier example, your belief that Brown owns a Ford or Brown is in Barcelona was lucky, and it could have easily been false, in Brown had happened to be somewhere else. A safety condition can also explain why the merely statistical evidence in the *Prisoners* case does not suffice for a conviction, and why it is not possible to know the lottery propositions, which makes it attractive as a necessary condition on legal proof. A belief that your lottery ticket will lose is not safe. In nearby worlds in which your ticket wins, the odds that it will lose are still low, and you believe on the basis of these odds that your ticket will lose. So, a belief that the ticket will lose could be easily false. Similarly, in the *Prisoners* case, a guilty verdict is not safe. In nearby worlds, the defendant is not guilty — he is the innocent prisoner who ran and hid during the riot – and the merely statistical evidence is the same.

Can the safety condition explain why the evidence is insufficient for a conviction in the *Traffic* and *Trading* cases? I think not. If there is a safety condition on knowledge, then the verdicts in the *Traffic* and *Trading* case are safe, because the evidence supports knowledge that the defendant is guilty. And the knowledge may be prior to safety. We figure out which worlds are nearby based on what we know, but not the other way around [Williamson, 2000]. After all, is it really the case that in nearby worlds, the innocent prisoner ran and hid instead of participating? Maybe we know that the innocent prisoner, whoever he is, is a pacifist, and it would be a remote possibility that he would join in.

So in order for the safety condition to provide an explanation of what goes wrong in these cases, safety must be necessary for legal proof, but not a necessary condition on knowledge (which means that this explanation occupies a very strange location in the conceptual space of epistemology.) Nevertheless, I think it provides an unsatisfying alternative explanation for the

cases. The evidence in the *Traffic* and *Trading* cases is safe. In the nearby worlds in which Annie speeds through the traffic light every day, she speeds through on June 17. If she hadn't sped through on June 17, it must have been because a strange coincidence happened: for example, she just so happened to be ill and remained home from work, which is a remote possibility. In the nearby worlds in which Bob says he is going to sell his stocks he does. If he hadn't sold his stocks, it must have been because something interrupted him: he had a last-minute change of heart, urgent business to attend to, or he couldn't reach his financial managers, all of which are remote possibilities. The safety condition will not do. What about sensitivity?

> **Sensitive Verdicts:** In the nearby worlds in which the defendant is innocent, the defendant is not convicted.

Like the safety condition, the sensitivity condition has been used to explain why you can't know lottery propositions, and why merely statistical evidence does not suffice for a conviction. The belief that a lottery ticket will lose is not sensitive. In the nearby worlds in which the ticket wins, the odds would lead you to form the same belief that the ticket will lose. A conviction in the *Prisoners* case is not sensitive. In the nearby world in which the innocent prisoner is on the stand, the merely statistical evidence would still lead the jury to deliver a guilty verdict.

Like the safety condition, the sensitivity condition is most often endorsed as a necessary condition on knowledge. But if sensitivity is to explain why the evidence does not support a conviction in the *Traffic* or *Trading* cases, it cannot be a necessary condition on knowledge, because the evidence supports knowledge in these cases. So this explanation would also occupy a strange location in conceptual space. I'd think that on the most plausible account of the sensitivity condition, the evidence in these cases is sensitive, and is evaluated using backtracking conditionals. In the nearby worlds in which Annie does not speed through the traffic light on June 17, she does not have a speeding habit, and she does not speed on the rest of the days. In the nearby worlds in which Bob does not trade the stocks, he also does not tell Carol that he would trade the stocks.

But to appeal to sensitivity would require treating the cases in a different way: In the *Traffic* case, in the nearby worlds where Annie does not speed through the traffic light on June 17, she still has a habit of speeding, but something unusual happened that day — maybe she become violently ill and did not attend work — and the courtroom evidence is the same. In the *Trading* case, in the nearby worlds where Bob does not sell his stocks, he had a last-minute change of heart, or urgent business to attend to, or couldn't get a hold of his financial managers, and in all of these cases, the evidence remains the same. So, the evidence in these cases is not sensitive.

So, the sensitivity account has some appeal, but I take this to be a virtue of my account, which can explain both the successes and the failures of the sensitivity condition. As the sensitivity condition was assessed in the following way: hold fixed all of the evidence (or the state of affairs) up until the time of the crime. Then suppose that the crime was not committed, and then assess whether the evidence would still support the conviction. This is very similar to the counterfactual test for causation [Lewis, 1973]. If C causes E, then if you hold fixed the state of affairs up to C, if C does not happen, then E will not happen. The counterfactual test has been taken to provide an account of causation, so my account of legal proof could explain why a sensitivity condition could be appealing. When the courtroom evidence fails to include trace evidence, it will probably fail the counterfactual test, and then fail to be sensitive.

My account of legal proof would predict the following: counterexamples to counterfactual accounts of causation will also be counterexamples to the sensitivity condition as a necessary condition for legal proof. And in fact, this is the case. Consider the following case:

> **Detective:** David is on trial for murder. The circumstantial evidence: the bloody knife with David's fingerprints and a diary entry where David confessed to the crime strongly supports David's guilt. The detective takes the stand and makes a shocking confession: if David had not committed the murder, she would have, and would have planted the evidence to look just the same.

A conviction in this case is not sensitive to David's guilt. If David had not committed the crime, the detective would have, and the evidence would still support a conviction. But a guilty verdict in this case is still perfectly appropriate. So, sensitivity is not necessary for legal proof. Here's another case that makes a similar point:

> **Tarot:** Ellen is on trial for murder. Her husband died of arsenic poisoning, and Ellen took out a life insurance policy against her husband, purchased the arsenic, and wrote in her diary that she wished her husband dead. She claims an alibi: she was with her sister Francis at their beachfront estate the weekend Ellen's husband died. Francis knows whether the alibi is true or not, and she has conflicted feelings about both her sister and the rule of law. She decides to leave the decision about whether to tell the truth to fate and draws tarot cards. If she draws a Seven of Spades (1/78 chance), she'll lie. If she draws any other card (77/78 chance), she'll tell the truth. She testifies that Ellen's alibi is true.

The evidence in this case suffices for a conviction. It's similar to any other case where there is circumstantial evidence against the defendant, and an eyewitness. Any time an eyewitness testifies, there is some chance that she will bear false witness, for any number of reasons (forgetfulness, intentional deception, and so on). In this case, the chances are made precise. A guilty verdict is not sensitive to Ellen's guilt. In a nearby world in which Ellen is innocent, her sister draws the Seven of Spades and does not corroborate her alibi, which would lead to a guilty verdict. Again, we have a case that has the correct causal structure, but lacks sensitivity, where the evidence suffices for a conviction.

1.4 A General Epistemic Thesis

If jurors hold out for trace evidence in order to deliver a guilty verdict, then jury verdicts will be less accurate. As we have seen, there are cases where it may be extremely probable that a defen-

dant is guilty but should not find the defendant guilty. Some find this conclusion unacceptable. Papineau [2021] writes: "We are being seduced by archaic ways of thinking into procedures that positively hinder our attempts to punish the guilty and save the innocent. We need to stop thinking in terms of knowledge," and Enoch, et. al [2021], characterize this as a kind of unacceptable "epistemic fetishism." But I think this line misses the point. Criminal justice encompasses more than convicting the guilty and acquitting the innocent. These are the just substantive outcomes of a jury trial, but justice also requires procedural due justice. Rawls [1971] illustrates the distinction between procedural and substantive justice with the example of a baseball coin flip that determines which team gets to start. Either substantive outcome that the Yankees start or the Red Sox start are both acceptable, but it would be unacceptable if the referee skipped the coin toss and picked the Yankees to start. The Red Sox would have a complaint on procedural grounds. Similarly, the outcomes of court proceedings matter, but so do the procedures that lead to them.

The requirement that a guilty verdict requires trace evidence is a matter of procedural justice. And in fact, this may run deep into the justification of criminal punishment itself. Kant writes: "the only time a criminal cannot complain that a wrong is done him is when he brings his misdeed back upon himself." We want punishment to be brought about by the crimes people have committed. That crimes must cause punishments is a matter of procedural justice, and one that places a requirement on the evidence a jury must have in order to deliver a guilty verdict. Legal proof requires trace evidence. But this leads us to a much more general epistemic point. All of our interpersonal reactions and reactive attitudes requires knowledge on the basis of trace evidence.

Here are some examples. A professor must read their student's term paper before assigning a course grade, even if they have had the student in class before and know that they would produce an excellent A+ term paper. An official at a high school swim meet must declare a winner based on what they see in the pool, even they knew in advance that the clear frontrunner (say, an Olympic prodigy) would win the race. You may know that your roommate will forget to water the plants while you are away— she is so scatterbrained and always forgets these things — but you can't

blame her until you get back home and see that the plants are wilting. Election officials must count the votes of an election before a candidate is sworn into office, even if everyone knows on the basis of opinion polls or other sociological evidence (one of the candidates is a popular Republican incumbent candidate in a deep red congressional district) that one candidate will win.

What do all of these cases have in common? Some of our actions are *predictions* and others are *reactions*. For example, bringing an umbrella on a raining day is predictive and not reactive. The decision to bring an umbrella should depend on whether it will rain, but it does not need to be a causal response to the rain. It could be a response to a weather report that says it will rain. All that matters is that we have an umbrella on a rainy day. Another set of actions and attitudes are *reactive*, they are actions in which people get what they deserve , on account of what they have done. This set includes our reactive attitudes (e.g., praise, blame, resentment, indignation, guilt), our evaluations (e.g., assigning grades or calling a race). Reactive attitudes are often held to be subject to a knowledge norm.[6] From Buchak [2014]: "Blame someone if and only if you believe (or know) that she transgressed and blame her in proportion to the expected severity of the transgression." The knowledge norm for blame can be underscored with the role that assertion plays in blame. When we blame our roommate for not watering the plants, we may assert: "you didn't water the plants!" And it sounds quasi-Moorean to say: "I don't know if you watered the plants or not, but I blame you for not watering them!"

These are all cases in which people should get what they deserve (an A+, a medal, blame, elected) on some basis. This basis should play a causal role in what happens. In general, it's important to secure the right causal link between what people deserve, and what they receive. It's important for crimes to bring about punishment, transgressions to bring about blame, merits to bring rewards, and to ensure this causal link, we must have evidence that is caused by or causally downstream from the actions of other people. Legal proof requires trace evidence. And

[6]Enoch & Spectre [2021] argue that it is wrong to hold reactive attitudes towards someone — including resentment and blame — on the basis of merely statistical evidence.

so do all of our reactions.

1.5 Epilogue: The Meno Problem

In the *Meno*, Plato raises a puzzle about the value of knowledge. Suppose you need directions to Larissa, and could ask one of two guides: one who knows the way, and the other who has a mere true belief. Either guide will point you in the right direction, and you will arrive at Larissa just the same. So why is knowledge valuable? Plato's solution comes in the form of a metaphor. The statues of Daedelus are valuable, but only if they are chained down so they cannot run away. The statues are tethered to the ground, and knowledge is tethered to the truth. Socrates says:

> "To acquire an untied work of Daedalus is not worth much, like acquiring a runaway slave, for it does not remain, but it is worth much if tied down, for his works are very beautiful. What am I thinking of when I say this? True opinions. For true opinions, as long as they remain are a fine thing and all they do is good, but they are not willing to remain long, and they escape from a man's mind, so that they are not worth much until one ties them down by giving an account of the reason why [*aitias logismos*]. And that, Meno, my friend, is recollection, as we previously agreed. After they are tied down, in the first place, knowledge is prized higher than correct opinion, and knowledge differs from correct opinion in being tied down. And that, my friend, is recollection, as we have previously agreed (*Meno* 97e-98a)."

On orthodox readings of this passage, *aitias logismos* entails justification, or the ability to provide reasons for one's belief. On another standard reading, what makes knowledge valuable is its *stability* [Williamson, 2000]. If one knows the way to Larissa, they are unlikely to change their mind. Even if the road looks different than they remember, or their companion tells them that they think it's the other way, they will still know which road leads to Larissa. And on yet

another reading, Socrates does not have in mind knowledge in the contemporary sense when he speaks of *episteme*, instead, he's interested in understanding, which includes explanations of the phenomena, and this is *aitias logismos* [Fine, 2004].

Sometimes, I wonder if Plato had something else in mind, if he too was interested in the causal origins of a belief: one translation of *aitias logismos* is "working out the cause" [Hyman, 2010]. And there may be, at least on a playful and anachronistic reading, more hints that *aitias logismos* is a causal notion. Socrates says:

> "it is through true opinion that statesman follow the right course for their cities. As regards to knowledge, they are no different from *soothsayers and prophets.* They too say many true things when inspired, but they have no knowledge of what they are saying [emphasis mine] (*Meno*, 99d)."

The choice of example is striking: the future-telling soothsayers and prophets — people who lack a causal connection to what they claim to know — are the paradigm case of people who have no knowledge of what they are saying. Socrates emphasizes *recollection* as a means to knowledge:

> "As the soul is immortal, has been born often and has seen all things here and in the underworld, there is nothing which it has not learned; so it is in no way surprising that it can recollect the things it knew before, both about virtue and other things (*Meno*, 81d)."

As we study mathematics or cultivate virtue, we do not learn anything new. Rather, we recollect what the soul has learned when it previous incarnations, perhaps even from the contact that the soul had with the forms themselves prior to embodiment. If this myth is to be believed, it would seem that even our knowledge of mathematical truths stands in a causal relationship to those truths.

If Plato intended to give a causal theory of knowledge, unfortunately, he would be mistaken. For reasons we have already seen, the causal theory of knowledge does not work (although the

Meno has a solution to one problem for the theory: mathematical knowledge.) Nevertheless, Plato identifies something of value: an epistemic state that is tethered to the truth.

2 | The Ethics of XAI

2.1 Introduction

Alice has a mole on her skin. She shows it to her dermatologist, who takes a picture of the spot and uploads it to a computer program, which classifies it as melanoma. The dermatologist looks at the computer output, and tells Alice she has melanoma. Alice asks: how can it tell? The dermatologist tells her that he has no idea how the algorithm works, but that it's accurate, and she should have the mole removed.

Bob is up for parole. He fills out a survey on topics that include his education and childhood, the criminal activity of his friends, and his moral views. His answers are run through an algorithm that assesses his risk for committing another crime upon release. He's denied parole, because his risk score is too high. How does the algorithm work? Bob has no clue — and neither do his lawyers. The statistical model that the algorithm uses is not public information.

Cindy writes a final essay for a college writing class. The professor runs it through a computer program, and it gives her a B. How does it work? She's not sure, but it usually gives students the same grades she would. Cindy gets an B.

The technology, in all of these cases, is real.[1] So is the *opacity*. The people whose lives they affect have no clue how the algorithms make their decisions. The algorithms map inputs

[1]See Chan et. al [2020] for an overview of AI applications in dermatology, Park [2019] on the use of the COMPAS algorithm for parole decisions, and Attali & Burstein [2006] on the development of the ETS e-rater algorithm for essay scoring.

to outputs — images to diagnostics, survey responses to risk scores, essays to grades — but it's unknown which features of the input the algorithm is sensitive to, or how they combine to arrive at the output.

Some of the opacity is due to *secrecy*. The COMPAS algorithm, which is used to make predictions about future criminal behavior of parole candidates in many US jurisdictions, is based on a statistical model that is a trade secret, and is not available to the public.

Some of the opacity is due to the *complexity* of the algorithm's statistical models. Often, when machine learning techniques have been used to construct the statistical models, the algorithm is a "Black Box," and it is an engineering challenge to figure out how it works.[2] The models are constructed using massive amounts of training data, for example, pictures of thousands of skin lesions along with a label that indicates whether it is cancerous or not. The algorithm then uses sophisticated mathematical techniques — like convolutional neural networks or high-dimensional regression — to learn how to classify the images. In the case of the melanoma diagnostic algorithm, the program picks up on patterns that do not map onto any of the concepts that dermatologists have. While a dermatologist may look at a spot and use a variant of the ABCDE mnemonic (asymmetrical, border, color, diameter, evolving), the algorithm is sensitive to intricate patterns of light and dark that seem random or incomprehensible to actual dermatologists. The result is an accurate algorithm — as accurate as dermascopy — that works in mysterious ways.

Another example. In the essay grading case, the algorithm is given training data, which consists of thousands of essays that have been graded by people, with an eye towards organization, style, word choice, vocabulary, and so on. The algorithm calculates the values of variables that correlate with the scores (e.g., the average word length for the quality of word choice, or the use of marker words like "because" for structure), and then uses a regression model to weight these variables to calculate the final score. The scores assigned by the algorithm approximate the scores

[2]For details about the engineering challenges, the successes and failures, and the DARPA initiative towards XAI, see Holzinger et. al [2018], Gunning et. al [2021], Gade et. al [2020], Kamath et. al [2021]

assigned by the human graders, but the students who submit essays, the educators who rely on the algorithms, and the engineers who created the algorithm, do not understand how it works.

Algorithmic opacity is a source of significant moral concern. It has been characterized as a "central issue" in the ethics of AI. Opaque algorithms are frequently described as "Kafkaesque," in reference to Kafka's *The Trial*, a short story in which a man is arrested by unidentified agents from an unidentified agency for an unspecified crime.[3] The threat of "algocracy" looms, and algorithmic opacity has created a fear that human decision-makers will be replaced with automated tyrants, who rule in a way that's beyond human understanding and that threatens the existence of our democratic processes.[4] The concern about opaque AI has prompted policy changes. The EU has established a "right to explanation" which entitles individuals to explanations for a range of algorithmic decisions that affect their lives, for example, in job hiring or mortgage lending [Kaminski, 2019]. The creation of explainable artificial intelligence (XAI) has become a scientific enterprise for engineers and computer scientists.

Some of the demand for XAI is to ensure the *accuracy* of the algorithms, and the judgments that are based on them. For example, if people understand how the algorithms work, they can more effectively pool their own information with the algorithmic outputs. For example, consider the case of the melanoma diagnostic algorithm. Some have picked up on the fact that spots circled in blue pen tend to be melanoma. This is a pattern in the training data: dermatologists tend to circle the spots that they are most worried about, and so the algorithm associates blue pen with cancer [Winkler et. al, 2019]. If the algorithm diagnoses a patient with melanoma just because their spot has been circled in blue pen, it is useful for the dermatologist to know this. If they know this, they should not take the algorithmic results to be additional evidence for the diagnosis of melanoma, and maybe they should erase the pen and upload a new photo.

Some of the demand for XAI is to discover and correct *bias* in the algorithms. For example,

[3]This analogy is drawn by Vrendenberg [2022], Moskvitch, [2013], Coeckelbergh, [2022], Selbst & Barocas [2018], Johnson [2014], Katsh & Rabinovich-Einy [2017], Wang [2018], Colaner [2022], Pasquale [2015], and many others.

[4]See Danaher [2016, 2020], Colaner [2022], Lukas et. al [2021], Aneesh [2006].

an algorithmic essay grader may evaluate the vocabulary of the essay writer by counting how many unique words they use, and how "common" or "advanced" the words are. However, which words are rare is culturally specific: "camel" may be a common word for Arabic speakers and an uncommon word for Japanese speakers, whereas "tofu" may be a common word for Japanese speakers and an uncommon word for Arabic speakers. If only "tofu" counts as an advanced word, then this creates a bias in favor of Japanese speakers and against Arabic speakers [Naismith, 2018]. If this lexicon is made available, then this bias could be discovered and corrected.

In this essay, I'll argue that XAI is important for reasons that extend beyond the bias or accuracy of the algorithms. Specifically, XAI is important to ensure that algorithms are acting for the right reasons. In a broad range of applications, the algorithms are used to give people what they deserve, on the basis of what they have done. They *respond* to what other people have done — they allocate punishments, rewards, and evaluations. Algorithms are used to referee sports games, grade student essays, issue speeding tickets and library late fees, evaluate teachers. These algorithms must act for the right reasons and with the right evidence. XAI is important to ensure that they do.

Section I will characterize this class of algorithms, and argue that it is is broader may initially meet the eye. One problem with the COMPAS algorithm for parole decisions is that it makes a *prediction* about the future conduct of a parole candidate, where it should be a *reaction* to the candidate's conduct in prison. The problem is not just that COMPAS is a "Black Box," it's that COMPAS a ""*crystal ball*", and it follows that it falls into this class of algorithms for which there is a special demand for XAI.

Section II will argue that these algorithmic responses are subject to epistemic norms. In Chapter 1, we saw that our responses are subject to epistemic norms. A similar epistemic norm holds for algorithmic responses. The algorithms must know that the basis for their response holds, and these responses must be caused by what other people have done. XAI is important to ensure the algorithms are following these epistemic norms.

Along the way, this essay will connect the dots between opacity and other moral concerns that arise in the ethics of AI literature: the use of proxy variables (for race, gender, and so on), the right to be treated as an individual, the right to appeal decisions, the threat of algocracy, and how to make sense of the epistemology of algorithms.

2.2 The Black Box and the Crystal Ball

2.2.1 Predictions and Responses

In some cases, algorithmic opacity seems like an urgent problem — the stuff of nightmares. Opacity in the realm of criminal punishment is especially unsettling, it's "Kafkaesque." Algorithmic criminal punishment still only exists in the realm of science fiction. (See: Minority Report, in which artificial intelligence systems predict and punish crimes before they happen.) Although artificial intelligence finds application in the law in parole hearings and predictive policing, and algorithms issue a range of punishments outside of criminal law, from speeding tickets to library fines.

In other cases, the opacity of algorithms doesn't seem like that big of a deal. Consider the melanoma diagnostic algorithms. According to some medical ethicists, the accuracy of a diagnostic tool is of primary importance, and it just doesn't matter that much whether we understand how it works [London, 2019]. It wouldn't be worth it to swap out an accurate opaque algorithm for a less accurate explainable algorithm. Medical practice is full of effective treatments that have unexplained causal mechanisms. Lithium is an effective mood stabilizer and aspirin has anti-inflammatory effects, but the medical community can only speculate as to why. These effective treatments are used instead of less effective treatments that have a well-understood mechanism of action.

Here's a morally significant difference between the cases. Criminal punishment should be a

response to a crime. It should give somebody what they deserve on the basis of what they have done. It's one of many responses that we may have to the actions of others, in which we allocate to them what they deserve, on the basis of what they have done. Claims about what people deserve are familiar:

- Abby is awarded a C on the basis of scoring a 70% on the quiz.
- Bill deserves punishment on the basis of committing a crime.
- Carol deserves a gold medal on the basis of swimming the fastest.

Algorithms are often tasked with responding to what people have done, and giving people what they deserve on the basis of what they have done. For example, algorithms are used to referee sports games, grade student essays, issue speeding tickets and library late fees, evaluate teachers, and so on.

A melanoma diagnosis is not one of these responses. A melanoma diagnosis is not a response that gives someone what they deserve on the basis of what they have done. Nobody a deserves a cancer diagnosis. A patient may be entitled to an accurate diagnosis, or owed an accurate diagnosis, or maybe even have a right to an accurate diagnosis, but a cancer diagnosis is not something that a patient earns through their actions or deserves on the basis of what they have done. The algorithm's output is not a reward, punishment, evaluation, or anything of the sort.

Responses are often associated with Strawsonian reactive attitudes, which "include resentment, gratitude, forgiveness, anger, or the sort of love for which two adults can sometimes be said to feel reciprocally, for eachother [Strawson, 1963]" Like punishments and rewards, evaluations, merits and demerits, reactive attitudes are held for some basis: you are grateful to someone for what they've done for you, and resentful to someone for what they have done to you. The award for an achievement has been characterized by Sidgwick as a kind of "gratitude universalized." The reactive attitudes in the participant stance stand in contrast with attitudes in the

objective stance: "To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained; perhaps simply to be avoided [Strawson, 1953]." To hold a reactive attitude is to engage with someone in the participant stance, as a member of the moral community who can be held responsible for their actions.

Reactive attitudes have fittingness conditions — they are only appropriate when held for the right reasons. It's only appropriate for you to blame your roommate for forgetting to water your plants if they actually forgot to water your plants. It's only appropriate for you to feel gratitude to your partner for taking out the trash if they actually took out the trash. Our responses have similar fitingness conditions. It's only appropriate to punish Bill if he actually committed the crime, and it's only appropriate to give Carol a gold medal if she actually swam the fastest. Of course, there could be good reason to punish Bill even if he didn't commit the crime (to send a message to people considering a life of crime), or good reason to award Carol a medal even if she didn't swim the fastest (she would be crushed if she lost), but there's still something that goes wrong in these cases — the reactions are inappropriate [Feinberg, 1970]. And in the case of criminal punishment, it's even worse than inappropriate, it's an injustice.

2.2.2 Algocracy and the Rule of Law

Opacity has raised concern about the dawn of the *algocracy*, or the "rule by algorithm." In this dystopian future (or present?), our democracy is replaced with the tyrannic rule of the algorithm, and we are subject to the whims of the machine. In the dark about how the algortihms will respond to what we do, we constantly toe an invisible line. From Danaher [2016]: "there is such a thing as the threat of algocracy. This is a threat to the legitimacy of public decision-making processes, which is posed by the opacity of certain algocratic governance systems."

Secret laws are taboo. The first written texts, in pictorial languages, include criminal legal

codes; the code of Urukagina in the Sumerian language is dated to 2600 BCE. Secret laws are widely regarded as a human rights violation in our contemporary context, and are banned in the constitutions of many countries.[5] For good reason. Secret laws threaten the rule of law. From Locke [1689]:

> "the ruling power ought to govern by declared and received laws, and not by extemporary dictates and undetermined resolutions: for then mankind will be in a far worse condition than in the state of nature…without having any measures set down which may guide and justify their actions: for all the power the government has, being only for the good of the society, as it ought not to be arbitrary and at pleasure, so it ought to be exercised by established and promulgated laws; that both the people may know their duty, and be safe and secure within the limits of the laws."

If the laws are not public, people are not able to conform their actions to it, and the law will not regulate and coordinate behavior. The criminal justice system is not automated yet. Criminal laws are public. But there's something in the spirit of this worry — that algorithmic opacity threatens the rule of law — that seems right. It's the algorithmic responses that raise concerns about algocracy. These responses — and the basis for desert claims — often lie in the rules that govern our institutions, practice, and norms. Often, it's important that these rules be public, so that people can conform their actions accordingly. A teacher may tell their students when the test will be, and which topics will be on the test, so that they study. This is good pedagogy — it encourages students learn the course material.

The Kafkaesque concerns are also related to the ability to appeal. The institutions that these responses stem from are often beaurocratic, and people have an interest in understanding how to navigate this bureaucracy. The focus of the paper here is on "backwards-looking" reasons for XAI: in order to ensure that the algorithms respond correctly to our actions. Vredenburgh [2021]

[5]UN Universal Declaration of Human Rights, and the constitutions of many countries (including the United States), explicitly ban *ex post facto* laws, laws that retroactively criminalize or punish conduct that was innocent when done. Secret laws violate the ban on *ex post facto* laws.

explains for the importance of XAI for "forward-looking" reasons, so we know how to avoid (or elicit) algorithmic responses, and to appeal decisions. This underscores the fact that XAI is especially important when the AI systems are *responses* — the these are the kinds of algorithmic decisions that could be appealed. You can appeal a grade, or a referee decision, or a guilty verdict. But you cannot appeal a cancer diagnosis. You could ask for a second opinion, or for your dermatologist to think more about the diagnosis, but you can't appeal a diagnosis. It's just not the kind of thing that you can appeal.

2.2.3 Example: COMPAS and Algorithmic Parole Decisions

Some of our algorithms make predictions about what people will do in the future (or what they would do or would have done in hypothetical scenarios), but should be reactions to what people have done in the past. Cases like this may be more common than you think.

An example. The COVID-19 pandemic led to the cancellation of A-level exams in the UK educational system. The missing grades were replaced with an estimation of what student grades would have been if they had taken the exams, based on the student's class ranking, schoolwide data, and the distribution of exam scores from the student's school in past years [Coughlin, 2020]. The algorithm is opaque and biased, but also should have never have been used in the first place. This is because grades should be a response to what students have actually done, not what they might have done if given the opportunity. Grades should be responses, not predictions.

Another example: the COMPAS algorithm for parole decisions. This section will explain why COMPAS errors in making predictions about the future crimes of parole candidates, instead of responding to the conduct of the parole candidate during incarceration.

The COMPAS algorithm is used to make parole decisions in many US jurisdictions. Parole candidates fill out a 137 word questionnaire, on topics that include their early family life, "criminal attitudes," history of substance use, the criminal activity of their friends and family, educational attainment, employment history, housing stability, and so on. The COMPAS algorithm uses the

questionnaire as an input, and then outputs a risk score, an assessment of how likely the parole candidate is to commit another crime upon release. The algorithm makes a prediction about what the candidate will do upon release.

The algorithm has come under fire for three reasons: the first is that it's not accurate. Studies have shown that it correctly predicts whether a parole candidate will commit a crime 63% of the time (the developers' own estimate: 68%), which is similar to the success rate of random research subjects recruited from the internet. Second, the algorithm is racially biased. It's more likely to incorrectly predict that black defendants will recidivate than white defendants. Finally, the algorithm is opaque. The statistical model that COMPAS uses is a trade secret, and not available to the public.

But there's another problem with COMPAS — a problem that extends beyond its inaccuracy, bias, and opacity. The problem with COMPAS is not just that it's a "Black Box." The problem is also that it is a *crystal ball.* The parole decisions based on COMPAS are predictions of what people will do in the future, instead of reactions to what people have done in the past. As a result, the subsequent parole decisions are made for the wrong reasons. They do not give people what they deserve on the basis of how they have done their time. This can be illustrated with the following cases (see Bell [2021] for a similar pair of cases):

> Paul is up for parole. He is in a medium security prison for bank robbery, which is his first criminal offense. He's a 28 year old man. He's had a tough upbringing — including poverty and housing instability. He dropped out of high school and several of his friends and family members have been in prison. In prison, he's expressed a commitment to "turn his life around." The prison guards have described him as a "model prisoner," he finished his GED, and worked his way up to head mechanic.

The COMPAS algorithm (and our best sociological models) may assign Paul a high risk score. His youth, gender, education level, and history of housing instability are all risk factors for recidivism.

Nevertheless, he seems like a strong candidate for parole. He's somebody who has done "all of the right things." He's earned it, and he deserves parole on the basis of his good time. Contrast Paul with Pat:

> Pat is up for parole. She is in a minimum security prison for bank robbery, which is her first criminal offense. She's a 65 year old affluent college-educated woman with a strong family support system. While incarcerated, she's expressed no remorse for her crime, has broken prison rules, was rude to the other prisoners, and refused to participate in any rehabilitative programs.

The COMPAS algorithm (and our best sociological models) may assign Pat a low risk score. Her age, socio-economic status, education, and strong family support system are all predictors of success outside of prison. Unlike Paul, she has not "done all of the right things." For the most part, she's stayed out of trouble, but hasn't done much else to deserve parole. Maybe she should be granted parole anyway, but given the choice between granting parole to Pat or granting parole to Paul, I think we should go with Paul.

To grant parole on the basis of a prediction of a candidate's future actions can have absurd consequences. The risk factors for future criminal actions include harsh treatment in prison, criminal actions of friends, and a history of childhood abuse. To deny someone parole for these reasons would be backwards, reek of "guilt by association," or be downright cruel. It's also at odds with the justification for parole in the first place. The parole system was created to to encourage "good time" for which a parole candidate is granted a "ticket of leave," an early release and a certificate that indicates that they have taken steps to successfully reenter society [Moran, 1945]. Bell [2021] argues that parole should be a credential, granted for time well spent. Like other credentials (a diploma or a license) it should granted on the basis of what people have done in the past, and not what people will do in the future. Parole should be granted on the basis of the candidate's conduct during incarceration, and not on the basis of their future criminal actions.

You may worry about this. Surely, you may think, there are people who should be kept in prison for the safety of people who are outside of prison. Imagine somebody who has done terrible crimes outside of prison, and has vowed to commit crimes again upon release. They have spent their time scheming, plotting, and threatening future crimes. Surely, this person is dangerous and should stay in prison. I agree.[6] But parole can be denied to this candidate on the basis of their conduct during incarceration. Their time in prison was not well spent. Time that was spent scheming, plotting, and threatening will not earn somebody parole.

A link will still exist between future crimes and parole decisions. Good time is spent doing things that contribute to rehabilitation, and that lower the chance that one will commit another crime. This includes participation in educational programs, substance use and mental health treatment, employment and vocational training, connecting with friends and family, and creating a plan for housing and employment post-release.

.

2.3 XAI and Epistemic Norms

Our reactions are subject to epistemic norms. Chapter 1 developed an account of the epistemic norms of reactions, using legal proof as the central example. In order to find a defendant guilty of a crime, a juror must know on the basis of trace evidence (i.e., evidence that is caused by or causally explained by the crime) that the defendant is guilty. The essay also suggested a general epistemic thesis: this epistemic norm holds for all of our responses.

Algorithms also need evidence in order to act. In order to respond appropriately to what others have done, the algorithms also need knowledge on the basis of trace evidence. The algorithm must be more than accurate — just as it's not enough for a jury to convict the guilty and acquit

[6] Actually, I think this is thorny. To hold someone in prison because of what they would do upon release resembles pre-punishment. Of course, there differences between denying someone parole and preemptive incarceration, there are also similarities.

the innocent — the algorithm must have the right evidence for the decisions it makes. XAI can help ensure that the algorithm comes to its decisions with the right evidence.

2.3.1 Knowledge and Merely Statistical Evidence

Punishment on the basis of merely statistical evidence is wrong. In order to convict a defendant of a crime, a jury must know that the defendant is guilty. Recall the following case from Chapter 1.

> **Prisoners:** 24 out of 25 prisoners who are incarcerated on a prison cell block form a riot and kill a guard. One of the prisoners does not join the riot and instead runs and hides. The riot is caught on a security camera, but when prosecutors review the tape, they cannot discern the identities of the defendants. They are all wearing identical jumpsuits, and the tape is too grainy to identify them from their faces. A prisoner, Bill, is selected at random and charged with murder.

The evidence in this case supports a high probability that Bill is guilty. Yet, the evidence does not support a guilty verdict. Why? Because the jury does not know that the defendant is guilty. A guilty verdict only just — only appropriate — if the jury knows that the defendant is guilty.

This knowledge norm generalizes to all of our responses:

- Give Abby a C only if you know she scored 70% on the quiz.
- Punish Bill only if you know he committed a crime.
- Award Carol a gold medal only if you know she swam the fastest.

Here are a few examples to illustrate these norms.

> **Quiz:** You have a stack of quizzes to grade. Your TA told you that he looked at all of the quizzes as the students turned them in, and nearly everyone in the class scored a

70%. 3 of the questions were tricky, and 7 were easy. A few excellent students scored 100%.

Suppose Abby's quiz is at the top of the stack of quizzes. You're tired and don't feel like grading. Should you just give Abby a C? No. You are very confident that she scored a 70% (as confident as you would be if you did not speak with your TA and graded the papers while this tired), but you don't know that she scored a 70%. She might have been one of the students who scored a 100%.

Swim: You are a referee for a swim meet. Carol is from a fancy high school, and the other swimmers are not. Swimmers from fancy high schools win swim meets 95% of the time. She has a beautiful dive into the pool. Unfortunately, you dozed off and didn't see who touched the wall first.

Should you just call the swim meet for Carol and award her the medal? No. You are confident that she won the race (as confident as you are in those cases where its hard to tell who touched the wall first), but you don't know that she swam the fastest. If you award her the medal, the other swimmers could complain: "you don't know that Carol swam the fastest!"

Just like us, algorithms must follow knowledge norms when they respond to what people have done. They must know that the basis for their response holds, and they cannot act on the basis of merely statistical evidence. Here are some examples. An automated library system could not count the number of books checked out and returned, notice that very few books were returned on time, and then issue fines to all of the borrowers. An essay grader could not grade all of the essays, and then give everyone a failing grade, because almost everyone has a failing grade. An algorithmic parole board could not look at the high rates of disciplinary infractions in a prison, and then use this disciplinary infraction deny parole to one of the the candidates from that prison.

These are silly examples. Few artificial intelligence systems work this way, because then they would not be intelligent systems. But some do. Teacher evaluation algorithms rely on schoolwide data — the progress of all of the students in a school — to assess the performance of individual

teachers for tenure, promotions, and merit pay. Some use of merely statistical evidence is okay. Even in the legal case, suppose that in the *Prisoners* case, the jury has both the videotape of the riot and a confession from the defendant. Then, a guilty verdict would be just. So the problem is not that the teacher evaluation algorithms use any schoolwide data at all. The problem would be if that the algorithm is not in a position to know about the performance of an individual teacher. So the devil is in the details — and this is the motivation for XAI. As long as the algorithm is opaque, it's difficult to figure out what the algorithm knows.

Another example. Consider the algorithmic essay grader. The algorithm uses a range of variables — word choice, essay length, average word length, the use of marker words like "because," "however," and "moreover," number of paragraphs, and so on — to come up with an essay score that approximates human graders. Many of these variables are measures (either directly or indirectly) of overall essay length, which the algorithm's developers take to be a feature, and not a bug. "Good writers have internalized the skills that give them better fluency ... enabling them to write more in a limited time [Winerip, 2012]" and the algortihm does approximate human graders well. Can the algorithm know what grades the students deserve — or whether their essays have a clear structure, sophisticated word choice, a thesis statement, grammatical sentences, and so on?

Again, it depends on the details. Suppose that Icelandic exchange students score lower on the essays than the rest of the population. The algorithm may pick up on the fact that essays that use certain words: like "Reykjavik," "swimming," "sheep," "volcano," "Inga," "wine," have low scores. If the algorithm detects this pattern, then it may start assigning low scores to any essay that uses these words — basically any essay that is written by an Icelandic exchange student. Can the algorithm know that an essay has a low score, just because it's written by an Icelandic exchange student who's used words that are familiar to them? I don't think so. This is an unacceptable use of merely statistical evidence. The algorithm's epistemic norm is not met. But it would be hard to detect this pattern without understanding how the algorithm works. If it really is the case that most Icelandic exchange student's don't do very well on the test, the algorithm may appear

accurate — assigning grades that are similar to what human graders would.

2.3.2 Causation and Proxy Variables

Our reactions must stand in the right causal relationship with what they react to. Sometimes we can know that the basis for our response holds, but the response is still not appropriate. Recall the following case of legal proof, from Chapter 1:

> **Traffic:** A traffic camera has been recording cars driving through a school zone for the past year. It has caught Annie speeding every weekday for the past year, except for one. On June 17, 2022, the camera batteries were dead, and they did not record the traffic.

The police can know, on the basis of this evidence, that Annie sped through the traffic light on June 17, 2022. They may lament: "we know she did it, but unfortunately, we didn't catch her." This generalizes to a wide range of cases. You need to grade Abby's quiz even if you know that she will get a C, because she always gets Cs. You need to watch the swim meet even if you know that Carol will win, because she's by far the best swimmer in the pool.

The causal basis for decisions is related to another problem, that arises in the ethics of AI: the use of proxy variables for protected groups to guide algorithmic decision-making. From Rai [2021]: "XAI techniques can be used to reveal whether attributes such as race or gender, or socio-economic and locational variables that proxy for them, are directly or indirectly used in black-box models so the models are biased against certain groups." A classic example. A bank is not allowed to use race as a factor in mortgage lending decisions — this would be illegal discrimination. However, an algorithm may be trained on a dataset that has the applications of people who have been granted loans, and whether or not they have defaulted on the loan. The algorithm may pick up on patterns in the applications, and find that people in some ZIP codes have defaulted on the loans. It may then start denying loans to people from that ZIP code. But in regions that are racially

segregated by ZIP code, ZIP code is a proxy for race. To deny loans to people who live in one ZIP code will be to deny loans to people from one racial group. Discrimination on the basis of ZIP code is racial discrimination.

Another classic example. A hiring algorithm may detect the fact that graduates of some colleges — like Mt. Holyoke College or Smith College — are not often selected for engineering jobs. As a result, it may not advance graduates from these colleges to the next round of consideration. But these are women's colleges, and to discriminate against graduates from women's colleges is gender discrimination. If the fact that these variables are used is not known, the algorithm could be covertly discriminating on the basis of race or gender.

When the algorithm relies on proxy variables, discrimination is a problem. In the case of algorithmic responses, the use of proxy variables can create an additional problem. If the algorithm relies entirely proxy variables, then the algorithmic responses may not be caused in the right way. For example, consider the algorithmic essay grader that identifies Icelandic exchange students, and then assigns them low essay scores. Even if the algorithmic grader accurate, and even if the algorithm knows that the Icelandic exchange students will not write good essays, they are still scoring the essays for the wrong reasons. The score should be guided, causally, by the merits of the essay.

2.3.3 Conclusion: Epistemology for Algorithms

At this point, you may worry: does the computer even have epistemic states? Does it even make sense to say that the computer knows that the student deserves an A, or the teacher deserves a raise, or that the parole candidate deserves parole? The computer doesn't have a robust set of concepts, or dispositions, or representational states, or many of the other mental states that may be prerequisites for knowledge. Maybe, you may worry, that talk of the algorithm's "knowledge"

is at best metaphorical, an anthropomorphism of the machine.[7] Let's just grant this metaphysics of the machine — suppose that the algorithms do not have these epistemic states. The algorithms just map inputs to outputs, with none of the epistemic states that show up in the epistemic norms. I think the explanation for the demand of XAI stands — at least as generating an epistemic requirement for the agents who employ the algorithms.

The professor knows that the TAs will assign the correct grades, and that the grades will be assigned on the right basis. If the TAs use merely statistical evidence, the professor cannot know that the grades will be assigned correctly. A professor uses an automated grading system. The professor knows that the computer will assign the correct grades, and that the grades will be assigned on the right basis. If the algorithms use merely statistical evidence, the professor cannot that the grades will be assigned correctly. So it could be that, at the end of the day, all that really matters are the epistemic states of people who are responsible for the algorithms. If so, we could still make sense of the "epistemic" states of the algorithm — this could be translated into the epistemic states of the people who engage with them.

Why XAI? XAI is important to ensure that algorithms act for the right reasons. In a broad range of applications, the algorithms give people what they deserve, on the basis of what they have done. They *respond* to what other people have done — they allocate punishments, rewards, and evaluations. These algorithms must act for the right reasons and with the right evidence. XAI helps us to ensure that they do.

[7] From Shevlin & Halina [2019]: "Currently, few people working in AI would literally attribute beliefs, thoughts, or feelings to machines."

3 | A Puzzle for Pragmatic Encroachment

Many epistemologists endorse the pragmatic encroachment thesis, including Unger [1984], Cohen [1988], Rysiew [2001], DeRose [1992, 2009], Hawthorne [2004], Stanley [2005], Hawthorne & Stanley [2008], Fantl & McGrath [2002, 2009], Weatherson [2011, 2012], Schroeder [2012], Ross & Schroeder [2014], Bloome-Tillman [2014], Kim [2016], and Moss [2018a, 2018b, 2021]. On this view, whether S knows that p depends on more than just epistemic features of her deliberative context — S's evidence, S's belief forming method, the truth of p, and so on. Instead, whether S knows that p also depends on the non-epistemic features of her deliberative context — her options and their stakes, and so on.

But this thesis, as standardly framed, is wrong. In section 1 of this paper, I'll consider cases that are used to motivate the pragmatic encroachment thesis and argue that they in fact create a problem for the thesis. Namely, there are equally plausible knowledge ascriptions, in these very same cases, that undermine the thesis. In section 2, I'll argue that the data can be accommodated in a contextualist account of knowledge and, formulate a knowledge norm that can account for the data. On this account, whether S knows that p depends on which of her options is salient to the knowledge ascriber.

3.1 A Puzzle For Pragmatic Encroachment

Here is a pair of cases that has been used to motivate the pragmatic encroachment thesis, from Ross Schroeder [2014]. [1]

> **Almond Butter — Low Stakes:** Sarah watches her roommate Hannah make three sandwiches: peanut butter, tuna, and almond butter. Hannah places the sandwiches in the fridge and tells Sarah: "the one on the right is almond butter and the one on the left is peanut butter" and then leaves the house. Sarah can't tell the nut butter sandwiches apart by sight, smell, or taste, but remembers what Hannah told her. She's babysitting a young child, Algernon, and needs to pack him a lunch. She knows he likes almond butter, tolerates tuna, and dislikes peanut butter.

What should Sarah do? She should pack the sandwich on the right. Why should she pack the sandwich on the right? The natural explanation: she knows that Algernon likes almond butter and that the sandwich on the right is almond butter. So, she should pack the sandwich on the right, the almond butter sandwich.

> **Almond Butter — High Stakes:** As in the Low-Stakes case, Sarah watches her roommate Hannah make three sandwiches: peanut butter, tuna, and almond butter. Hannah places the sandwiches in the fridge and tells Sarah: "the one on the right is almond butter and the one on the left is peanut butter" and then leaves the house. As before, Sarah can't tell the nut butter sandwiches apart by sight, smell, or taste, but remembers what Hannah told her. She's babysitting a young child, Algernon, and needs to pack him a lunch, and she knows he likes almond butter and tolerates tuna. However, she also knows that he has a severe peanut allergy and giving him a peanut butter sandwich would have catastrophic consequences.

[1] Similar cases appear in DeRose [1992], Cohen [1999], Stanley [2005], Brown [2008], and Fantl & McGrath [2009].

What should Sarah do? She should pack the tuna sandwich. Why should she pack the tuna sandwich? The natural explanation: she knows that Algernon has a severe peanut allergy and giving him a peanut butter sandwich would kill him. She should play it safe and give him the tuna sandwich. Though she takes it to be likely that the sandwich on the right is almond butter, she does not know the sandwich on the right is almond butter. If she gives him the sandwich on the right, the almond butter sandwich, she could be blamed: "you didn't know that was an almond butter sandwich!" Hawthorne & Stanley [2008] defend this role of knowledge ascriptions in blame for negligent behavior: "If a parent allows a child to play near a dog and does not know whether the dog would bite the child, and if a doctor uses a needle that he did not know to be safe, then they are *prima facie* negligent."

The High–Stakes and Low-Stakes cases are identical with respect to the *epistemic* features of Sarah's deliberative context, but differ with respect to the *practical* stakes, and whether Sarah knows that the sandwich on the right is almond butter and the one on the left is peanut butter. Hence the pragmatic encroachment thesis: the practical stakes make a difference to what Sarah is in a position to know.

Here's the problem for pragmatic encroachment, and a central point of this paper: In the High-Stakes case, there is a strong argument that Sarah does know that sandwich on the right is almond butter and the sandwich on the left is peanut butter. Suppose Sarah is looking at the three sandwiches and deliberating about which one to pack for Algernon. She picks up the sandwich on the left and wonders: "should I pack this for Algernon?" She should not. Why not? A natural explanation: she knows it is a peanut butter sandwich, and Algernon is allergic to peanut butter.

Suppose she gives the sandwich on the left to Algernon (maybe she's lazy and it was closer in reach, or she hates her nephew and wants him dead.) Has she done something wrong? Yes. Why? A natural explanation: she knew she was giving him a peanut butter sandwich.

It would be absurd for her defend herself with the claim: "I didn't know that was a peanut butter sandwich!" And we could reply: "Yes you did. You knew this was a peanut butter sandwich

and that Algernon is allergic to peanut butter. You should not have given him this sandwich that you knew he was allergic to."

Some of the absurdity of Sarah's claim could be explained away by pragmatic considerations. It sounds like she's trying to offer an excuse, and her action is not excusable. She should not have packed that sandwich for Algernon. Nevertheless, her claim "I didn't know that was a peanut butter sandwich" is false, and we can respond, truthfully, "yes you did, you knew that was a peanut butter sandwich!"

Well, you might say in response, there is an alternative explanation for why Sarah's action was wrong: she knew the sandwich was probably peanut butter, and Algernon is allergic to peanut butter. While it sounds right to say that she knew it was peanut butter, she in fact only knew that it was probably peanut butter, and this is enough to make her action wrong.

However, this response creates trouble for the pragmatic encroachment thesis. A similar alternative explanation could be given in the Almond Butter — Low Stakes case. Recall that in this case, Sarah should give Algernon the sandwich on the right, the almond butter sandwich. The natural explanation for this (which led to the pragmatic encroachment thesis) is that she knows it's almond butter. The similar alternative explanation: Sarah knew the sandwich was probably almond butter, and Algernon likes almond butter, and she should give him a sandwich that he will probably like.

3.2 A Solution For Pragmatic Encroachment

Many epistemologists who endorse the pragmatic encroachment thesis are subject-sensitive invariantists, including Ryesview [2001], Hawthorne [2004], Stanley [2005], Fantl & McGrath [2002, 2009], Weatherson [2011, 2012], Ross & Schroeder [2014], and Kim [2016]. According to the subject-sensitive invariantists, whether S knows that p depends on the practical features of S's deliberative context, but the meaning of the word "know" and the truth conditions of "S knows

that p" remain the same in every context that the knowledge ascriptions are made. The relationship between knowledge and the deliberative context are characterized by norms that link knowledge to practical reason, rational preference, or rational action, like the following [Hawthorne & Stanley, 2008]:

> **Knowledge Norm of Action:** S knows that p only if it is rational for her to act as if p.

In the Low-Stakes case, it is rational for her to act as if the sandwich on the right is almond butter and to give her nephew the sandwich on the right, so she satisfies the necessary condition on knowledge given by the knowledge norm. In the High-Stakes case, it is not rational for her to act as if the sandwich on the right is almond butter and to give her nephew the sandwich on the right, so she does not know that the sandwich on the right is almond butter.

The principle is unable to account for the fact that Sarah can know that the sandwich on the left is peanut butter. As we have seen, if Sarah goes ahead and gives Algernon a peanut butter sandwich, she could be blamed: "you knew that was a peanut butter sandwich!" Sarah knows that one nut butter sandwich is peanut butter, and the other is almond butter. So, if she knows that the sandwich on the left is peanut butter, she could also know that the sandwich on the right is almond butter, which is ruled out by the knowledge norm.

Of course, which actions are rational may depend on Sarah's goals. If Evil Sarah hates her nephew and wants to kill him, then it is rational for her to act as if the sandwich on the left is peanut butter and pack him that sandwich, and it is consistent with the knowledge norm that she knows the sandwich on the left is peanut butter. However, consider Lazy Sarah, who loves her nephew and wants him to live, but also wants to give him the sandwich on the left because she is lazy and it is closer in reach. If Lazy Sarah gives him the sandwich on the left, we can blame her: "you knew that was a peanut butter sandwich!" So an appeal to Sarah's motives will not solve the problem for the subject-sensitive invariantists.

Many other epistemologists who endorse the pragmatic encroachment thesis are contextualists about knowledge, including Unger [1984], Cohen [1988], DeRose [1992, 2009], Lewis [1996], Bloome-Tillman [2014], and Moss [2018]. According to the contextualists, the meaning of "know" and the truth conditions of knowledge claims vary depending on the context in which the claims are made. Some contextualists appeal to relevant alternatives: S knows that p only if S has ruled out every relevant alternative to p. Which alternatives are relevant? Salient alternatives may be relevant, as may alternatives with a sufficiently high probability, hence the pragmatic encroachment thesis.

From Lewis [1996]: "how high is sufficiently high? That may depend on how much is at stake. When error would be especially disastrous, few possibilities may be properly ignored." In Low-Stakes, not much is at stake in which sandwich Sarah packs for Algernon. So the possibility that the sandwich on the right is peanut butter can be ruled out, and Sarah knows the sandwich on the right is almond butter. In High-Stakes, much is at stake in which sandwich Sarah packs for Algernon. So the possibility that the sandwich on the right is peanut butter cannot be ruled out, and Sarah is not in a position to know that the sandwich is peanut butter. This is on the right track. But what the contextualist needs is an account of how the stakes interact with the salient options to predict that the claim that Sarah knows the sandwich on the left is peanut butter can be true.

Here's a contextualist picture that I think accounts for this data. Two parts. First, standard decision theory: what Sarah should do, in Low-Stakes and High-Stakes, is determined by how severe Algernon's allergy is, and the probability that sandwich on the left is peanut butter and the sandwich on the right is almond butter. In the Low-Stakes case, she should pack the sandwich on the right, the almond butter sandwich, because it is the option with the highest expected utility. In the High-Stakes case, she should not pack the almond butter sandwich, because it does not have the highest expected utility. And she should not pack the sandwich on the left, because it has the lowest expected utility.

Second, contextualism: the natural explanations for why Sarah should or should not take these options, from the previous section, summarize these facts in knowledge-based terms. In the Low-Stakes case, the explanation for why Sarah should pack Algernon the sandwich on the right appeals to the strength of her epistemic position. She is confident enough that it is almond butter to take that option: the difference between her confidence (e.g., her credence or epistemic probability) and absolute certainty that it is almond butter makes no difference to whether she should pack him. In the High-Stakes case, the explanation for why Sarah should not pack Algernon the sandwich on the right appeals to the weakness of her epistemic position. She is not confident enough that it is almond butter to take that option: the difference between her confidence and absolute certainty does make a difference to whether she should pack him the almond butter sandwich. The explanation for why Sarah should not pack the sandwich on the left appeals to the strength of her epistemic position. She is confident enough that it is peanut butter to rule out that option: the difference between her confidence and absolute certainty makes no difference to whether she should pack him the sandwich on the right. This suggests the following general principle:

> **Contextualist Knowledge Norm:** If an option ϕ is salient, S knows that p only if the difference between S's confidence in p and full certainty makes no difference to whether she should ϕ.

This account gets the data right. As we consider Sarah's options, and reason about what she should or should not do, the salience of the options change, and this changes which claims about what Sarah "knows" are true. Should she give Algernon the sandwich on the right? No, because she does not know it's almond butter. Should she give Algernon the sandwich on the left? No, because she knows it's peanut butter. If she kills Algernon with the peanut butter sandwich, this option is salient as we blame her: "you should not have done that — you knew that was a peanut butter sandwich!"

4 | How To Set Statistical Significance Levels

Many epistemologists endorse the pragmatic encroachment thesis: that whether S knows that p depends on more than just the epistemic features of her deliberative context — S's evidence, S's belief forming method, the truth of p, and so on. Instead, whether S knows that p also depends on the non-epistemic features of her deliberative context — her options and their stakes, and so on. I'm one of these epistemologists — Chapter 3 developed an account of the pragmatic encroachment thesis.

Many philosophers of science, scientists, and statisticians endorse a parallel thesis in science: non-epistemic considerations play a role in how scientists should formulate statistical significance tests. According to this thesis, significance levels should vary with the practical costs of errors. The standard formulation of the thesis is as follows.[1]

Standard Account:

- If a Type I error is costly, lower the significance level. That is, if it's costly to reject a null hypothesis if true, lower the probability that you will reject the null hypothesis conditional on the null hypothesis.

[1] A defense of the standard account can be found by philosophers in Rudner [1953], Hempel [1965], Douglas [2000], scientists in Brown [1983], Cranor [1993], Banerjee [2009], Liberman et.al [2009], statisticians include Neyman-Pearson [1933], Holland & Ordoukhani [1989], Gordon et. al [2009].

- If a Type II error is costly, raise the statistical power. That is, if it's costly to fail to reject a null hypothesis if false, lower the probability that you will reject the null hypothesis conditional on the the alternative hypothesis.

Neyman & Pearson incorporated the Standard Account into the development of their paradigm for statistical hypothesis testing:

> "If we reject H_0, we may reject it when it is true; if we accept H_0, we may be accepting it when it is false, that is to say, when really some alternative is true. These two sources of error can rarely be eliminated completely; in some cases, it will be more important to avoid the first, in others the second ... The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator [Neyman & Pearson, 1933]"

Philosophers of science who reject the value-free"ideal often take statistical significance testing to be a paradigm case of the role of non-epistemic values in science. The cost of errors should influence the evaluation of scientific hypotheses, and how significance levels are set. Rudner argues that the "scientist qua scientist" makes value judgements:"

> "But if this is so then clearly the scientist as scientist does make value judgments. For, since no scientific hypothesis is ever completely verified, in accepting a hypothesis the scientist must make the decision that the evidence is sufficiently strong or that the probability is sufficiently high to warrant the acceptance of the hypothesis. Obviously our decision regarding the evidence and respecting how strong is "strong enough", is going to be a function of the importance, in the typically ethical sense, of making a mistake in accepting or rejecting the hypothesis [Rudner, 1953]"[2]

[2]Rudner is concerned with significance testing in the frequentist paradigm: "I have obviously used the term "probability" up to this point in a quite loose and pre-analytic sense. But my point can be given a more rigorous formulation in terms of a description of the process of making statistical inference and of the acceptance or rejection of hypotheses in statistics. As is well known, the acceptance or rejection of such a hypothesis presupposes that a certain level of significance or level of confidence or critical region be selected."

The feminist empiricist literature characterizes the conclusions drawn from significance testing as an "inductive risk." In conducting significance test, one risks a Type I and Type II error. The researcher's non-epistemic values determine the acceptable risk of each error, and consequently, the significance levels.

> "The deliberate choice of a level of statistical significance requires that one consider which kind of errors one is willing to tolerate. ... If one wishes to avoid more false negatives and one is willing to accept more false positives, one should lower the standard for statistical significance. If one wishes, on the other hand, to avoid false positives more, one should raise the standard for statistical significance ... In setting the standard for statistical significance, one must decide what balance between false positives and false negatives is optimal. In making this decision, one ought to consider the consequences of the false positives and false negatives, both epistemic and non-epistemic [Douglas, 2000]."

The classic example: suppose drug manufacturers and regulators are evaluating a new therapeutic to determine whether it is safe. They test the null hypothesis that the drug is safe. If they reject the hypothesis that the drug is safe, they will not approve the drug. If they fail to reject the hypothesis that the drug is safe, they will approve the drug. There are two errors they could make: they could not approve a safe drug (Type I error), or they could approve an unsafe drug (Type II error). The Type I and Type II error rates trade off. If they lower the Type I error rate, they raise the Type II error rate. If they lower the Type II error rate, they raise the Type I error rate. According to the Standard Account, they should set these error rates based on the costs of the errors. If approving an unsafe drug is the more costly error, they should raise the Type I error rate and lower the Type II error rate, and if not approving a safe drug is the more costly error, they should lower the Type I error rate and raise the Type II error rate.

In this paper, I'll argue that non-epistemic considerations should play a role in statistical significance testing — but not in the way suggested by the Standard Account. Section I will

provide an overview of the Fisherian and Neyman-Pearson paradigms for significance testing, along with the alternative Bayesian paradigm.

In Section II, I'll argue that the Bayesian paradigm provides a better *theory* of scientific rationality than the frequentist paradigm, but that it does not follow that we should "abandon statistical significance" testing or "ban the p-value." Frequentist statistics can still be useful in *practice.* Bayesian researchers face cognitive and spatiotemporal limitations, and they need statistics in order to effectively condense, communicate, and learn from data. This is an important role for statistics, and one that significance tests play well. In some paradigm cases, if Bayesians conditionalize on the results of significance tests their posterior credences will approximate (and have similar expected inaccuracy as) the posterior credences that would result from conditionalization on the entire dataset. To run a statistical test is to condense the data into one-bit of information — it is to seek the answer to one "yes/no" question about the data. It's a mode of inquiry, a practice in the context of discovery, a context of science that is unquestionably value-laden.

In Section III, I'll formulate an account of the role that non-epistemic considerations should play in the construction of a statistical test: an account of how to set significance levels. The significance levels determine which Bayesian posteriors could result from conditionalizing on the outcome of the test. The value of a posterior credence distribution depends on non-epistemic values. To run a statistical test is to ask a question about a data, and the value of asking a question depends on the value of the answers.

In Section IV, I'll apply this framework to two cases. First, the classic example: how the FDA should set significance levels as they evaluate drugs for "substantial evidence of efficacy" and "proof of safety." Second, the application for which Neyman-Pearson significance tests were developed: the eugenicist project of searching for differences in heritable traits among ethnic groups, and in contemporary "neurosexist" and "scientifically racist" research projects. In the first application, the Standard Account gets things right. In the second application, the Standard Account errs, in ways that could have grave consequences.

4.1 Statistical Significance Testing

This section will provide an overview of two paradigms of statistical significance testing: Fisher's exact test and the Neyman-Pearson hypothesis test, along with an overview of the Bayesian approach to scientific inference.

4.1.1 Fisher's Exact Test

Fisher [1935] motivates his "exact test" or null hypothesis significance test (NHST) with a charming example of the lady tasting tea. Fisher invited a friend, Lady Bristol, over for tea. She took one sip of the tea, and complained: this tea was poured milk first, and not tea first. Fisher is incredulous: can she really tell whether tea has been poured milk first or tea first? He forms the hypothesis that she can only guess at random. If you place a milk first cup and tea first cup in front of her, and ask her to guess which is which, she has a 50/50 chance of guessing correctly. Fisher tests this hypothesis. He gives Lady Bristol eight cups of tea, four that were poured milk first and four that were poured tea first. She guesses all eight correctly. Fisher is flabbergasted. What is the chance that she would do this well, if she can only guess at random? Fisher calculates: 1.4% A small chance. So he rejects the hypothesis that the lady was guessing at random.

Here are the mathematical details of the exact test. For the sake of simplicity, I'll use a different running example than Fisher, with a simpler statistical model. Suppose that we have a coin and want to learn its bias. The coin flips are represented by a statistical model, P_θ. The statistical model specifies a probability distribution that has a parameter θ. A coin flip is represented as a Bernoulli random variable from P_θ: each flip has a θ chance of landing Heads and a $1 - \theta$ chance of landing Tails. The coin flips are modeled as independent and identically distributed random variables, and we will flip the coin n times in order to learn about its bias.

1. Select a null hypothesis

The first step is to formulate a null hypothesis. In Fisher's experiment, his null hypothesis was that Lady Bristol could tell the difference tea first and milk first cups of tea. For our experiment, let's consider the null hypothesis that the coin is fair. $H_0 : \theta = .5$.

According to Fisher, the null hypothesis must be simple and pointwise: it must specify an exact value of θ. Furthermore, the null hypothesis should posit a symmetry or a similarity. Examples: a coin is fair, a drug is equally effective as a placebo, or the Lady guesses at random.

2. Run an experiment

Let's run the experiment. Let's flip the coin $n = 30$ times:

H T T H H T H T H H T T T T H T H H H T H H H H T H H H T H

Call the entire sequence X, and label outcome of each coin flip X_1 to X_{30}.

3. Calculate the value of a test statistic

Next, calculate the value of a *test statistic* $T(X)$, a function of the data. In the Lady Tasting Tea experiment, Fisher counted the number of cups that the lady guessed correctly. For our experiment, let $T(X)$ be the number of Heads, so $T(X) = 18$.

3. Calculate a p-value.

Definition. A p-value is the probability, conditional on the null hypothesis, of observing a value of the test statistic that is *at least as extreme* as the one in fact observed.

For us, the p-value is the probability, conditional on the null hypothesis that the coin is fair, that the coin lands Heads at least 18 times. So $p = P_{.5}[T(X) \geq 18] = .18$.

4. Reject or do not reject the null hypothesis.

Finally, *reject* or *fail to reject* the null hypothesis. If the p-value is less than the significance level (i.e., α-level), reject the null hypothesis. The result is *statistically significant*. If the p-value is not less than the α-level, do not reject the null hypothesis. The result is not statistically significant.

How to set the α-level? Fisher does not say. In practice, the α-level is often set according to conventions that vary by discipline. The significance level is often .05 in the behavioral or social sciences, .01 in the biological and medical sciences, and less than .001 in the computational and physical sciences.

In our coin flip experiment, the p-value is .18, so the result is not statistically significant.

4.1.2 Neyman-Pearson Testing

The Fisherian paradigm has bugs that Neyman and Pearson's paradigm fixes [1933]. Specifically, the results of a Fisherian significance test depend on choices that are left to a researcher: the choice of test statistic, which outcomes of an experiment are "at least as extreme" as the ones observed, and how to set significance levels. This gives rise to arbitrariness concerns. The Neyman-Pearson paradigm makes principled recommendations about these choice-points.

1. Select a null hypothesis and an alternative hypothesis

As in the Fisherian method, select a null hypothesis. In our case, let's select $H_0 : \theta = .5$, the null hypothesis is that the coin is fair. In the Neyman-Pearson paradigm, we also select an *alternative hypothesis* H_1 to test against H_0. The alternative hypothesis can be pointwise and specify a single value, or can be composite, and range over several values. Let's let $H_1 : \theta > .5$, which so the alternative hypothesis is that the coin is biased towards Heads.

2. Construct a statistical test

To construct a statistical test, first select a *rejection region* R, a set of possible outcomes of the experiment. If the experimental outcome falls into the rejection region, *reject* the null hypothesis (and accept the alternative hypothesis). If the experimental outcome falls outside of the rejection region, *accept* the null hypothesis (and reject the alternative hypothesis). That is, if $X \in R$, *reject* the null hypothesis. If $X \notin R$, *accept* the null hypothesis.

How to select the rejection region? According to the Neyman-Pearson paradigm, researchers should construct a test that minimizes α, the probability of rejecting the null hypothesis, if the null hypothesis is true (Type I error); and β, the probability of accepting the null hypothesis, if the alternative is true (Type II error).

That is, the goal is to minimize the following error rates:

$$\alpha = P_{\theta=\theta_0}[X \in R]$$

$$\beta = \sup P_{\theta \in H_1}[X \notin R]$$

As we have seen, Neyman and Pearson argue that researchers should set these error rates based on the relative costs of errors: if the Type I error is costly, set α low; if the Type II error is costly, set β low.

In actual scientific practice, scientists often use an α-level that is set by convention (e.g., .05), and then select a powerful test at that level (i.e., the test that minimizes *beta*). For some distributions and set of hypotheses, there exists a *uniformly most powerful test*, a test that is the most powerful at some level if θ is any of the values in the alternative hypothesis.

In the case of our coin flip experiment, there is a uniformly most powerful test at the .05 level: $R = \{X | T(X) \geq 20\}$, where $T(X)$ is the number of Heads.

2. Accept or reject the null hypothesis

If the observed data falls in the rejection region, *reject* the null hypothesis. The result is statistically significant. If the observed data falls outside of the rejection region, *accept* the null hypothesis. The result is not statistically significant. Note that in the Fisherian exact test, a scientist can either reject or fail to reject a null hypothesis, but Neyman-Pearson introduces the *acceptance* of a null or alternative hypothesis.

In our experiment, the data does not fall in the rejection region. The coin landed Heads 18 times, and would need to land Heads at least 20 Heads in order to reject the null hypothesis that the coin is fair. The result is not statistically significant, and we should not reject the null hypothesis.

A definition. A p-value is the lowest α-level for which the null hypothesis would be rejected.

This is equivalent to the Fisherian definition, and in our experiment, the p-value is .18.

4.1.3 The Bayesian Approach

The Bayesian introduces a subjective credence function into scientific inference. The Bayesian starts with a prior credence function c over the value of the parameter θ in the statistical model, and then updates this credence function by conditionalizaton on the results of the experiment. The result is an updated credence function c_X. Formally[3]:

$$c_x(\theta) = c(\theta|X) = \frac{P_\theta(X)c(\theta)}{\int_{\theta_j} P_{\theta_j}(X)c(\theta_j)}.$$

To illustrate, in our running coin flip example, let's say we have two hypotheses: $H_0 : \theta = .5$ and $H_1 : \theta = .8$, and a prior credence function c over these hypotheses, such that $c(.5) = .5$ and $c(.5) = .8$. After we flip the coin and observe the sequence of coin flips X, we update our

[3]For this definition, c is a probability density function, but a corresponding definition could be given when c is a probability mass function.

credences by conditionalization, which results in the the new credence function c_x, where $c_x(.5) = c(\theta_0|X) = .44$ and $c_x(.8) = c(\theta_1|X) = .56$.

Another definition. Consider a case where the two hypotheses under consideration are pointwise, so $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. Then,

$$\frac{c(\theta_0|X)}{c(\theta_1|X)} = \frac{c(X|\theta_0)}{c(X|\theta_1)} \cdot \frac{c(\theta_0)}{c(\theta_1)}$$

.

The middle term, $K = \frac{c(X|\theta_0)}{c(X|\theta_1)}$ is called the "Bayes factor", and it quantifies how strong the evidence is for one hypothesis over another. It has all the information you need in order to figure out how to update your credences upon learning X.

If $K > 10$ or $K < .1$, the evidence is said to be "strong," and a Bayes factor threshold test, which indicates whether or not the data is strong evidence against the null hypothesis, has been suggested as an alternative to frequentist significance testing [Wakefield, 2009]. The frequentist can calculate a Bayes factor with the mathematical tools that they have. The calculation of a Bayes factor does not depend on a prior credence function, as $c(X|\theta) = P_\theta[X]$.

In our running example, the Bayes factor for our experiment is

$$K = \frac{c(X|\theta_0)}{c(X|\theta_1)} = \frac{.028}{.035} = .8$$

so the data is evidence (but not strong evidence) against the null hypothesis that the coin is fair.

4.2 Statistics in the Context of Discovery

4.2.1 The Problem With Statistical Significance

Here's the problem with statistical significance testing as a theory of scientific rationality — and the reason to favor Bayesianism — in a nutshell. The statistical significance test does not make use of a scientist's *total evidence* about a hypothesis. As a result, if a scientist's credences, beliefs, or actions are determined entirely by the results of a significance test, then significance testing will lead to irrational credences, beliefs, or actions.

For example, running a significance test could lead you to reject a hypothesis that — on the basis of your total evidence — you know to be true. For example, suppose that you know that a coin is fair. It's a normal coin, it's symmetrical, it's identical to other fair coins, and an Oracle and a physicist told you that it's a fair coin. You flip it 30 times and it lands Heads 20 times. You calculate the p-value: 0.049, and because this result is statistically significant at the .05 level, you reject the hypothesis that the coin is fair. If you consider your total evidence, you should not have a high credence that the coin is biased, or believe that the coin is biased, or act as if it's biased. So statistical significance testing cannot provide a correct theory of rational credence, belief, or action. This raises a question for which the frequentist has no good answer: if significance testing is not theory of credence, belief, or action, what exactly, is it supposed to be a theory of?

According to Fisher, if a null hypothesis is rejected, then it has been *falsified*. That is, the experimental results are "inconsistent" with the null hypothesis, and the data is "proof" that the hypothesis is false. But this is wrong. A coin landing Heads on 20 out of 30 flips (or even 30 out of 30 flips) is consistent with the hypothesis that the coin is fair. It's unlikely, but possible, for a fair coin to land Heads 20 out of 30 times. Fisher suggests that there is a role for prior information in the evaluation of a hypothesis, as he writes:

> "It is open to the experimenter to be more or less exacting in respect of the small-

> ness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. It is obvious that an experiment would be useless of which no possible result would satisfy him [Fisher, 1933]."

So Fisher must recognize that prior information — a researcher's total evidence — should (or does) play a role in the evaluation of a scientific hypotheses, but he gives no systematic account of how to use this information to draw conclusions.

According to Neyman and Pearson, to accept a hypothesis is to act as if it is true, and to reject a hypothesis is to act as if it false, the "rule of behavior." But this also will not do. It is often irrational to act as if a hypothesis is true, even if a significance test supports its acceptance. Likewise, it is often irrational to act as if that a hypothesis is false, even if a significance test supports its rejection. If you flip a coin that you know is fair 30 times, and it lands Heads 20 times, you shouldn't go ahead and act as if the coin is biased towards Heads. If you are a referee for a soccer game, and you need to flip a fair coin to start the game, this coin will work just fine. You shouldn't open your wallet or check your pockets in search of a fair coin. The American Statistical Association issued a warning against the "rule of behavior" [2016]: "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold."

Note that this behaviorist interpretation of frequentist statistics — in which accepting or rejecting a hypothesis is associated with an acting as if it is true or false — is the interpretation that the proponents of the Standard Account of how to set significance levels hold. They associate the cost of a Type I error with the cost of acting as if the null hypothesis is false (when true), and associate the cost of a Type II error with the cost of acting as if the null hypothesis is true (when false).

Some philosophers (and increasingly, the scientific community) have declared that it is time to "abandon statistical significance" [McShane, et. al, 2019]. The Bayesian epistemologists Howson and Urbach [1993] have strong words: "classical methods are set altogether on the wrong lines,

and are based on ideas inimical to scientific method." I agree that frequentism fails to give a theory of rational credence, belief, or action. I also agree with Kukla [2015] that accepting or rejecting a hypothesis has no straightforward relationship to rational credence, justified belief, knowledge, or the other epistemic states of interest in epistemology, and instead, rejecting a hypothesis has "various institutional consequences and preconditions ... it is necessarily embedded in a broad and rich social network of scientific practices and agendas." Frequentist statistics play an important role in scientific practice. The next section will develop an account of this role, along with an argument that frequentist statistics plays it well.

4.2.2 Statistical Significance in the Context of Discovery

A set of dichotomies runs through the traditional understanding of statistical hypothesis testing:

> theory | justification | inferential | ampliative | objective | value-free
>
> practice | discovery | descriptive | non-ampliative | subjective | value-laden

Statistical significance testing is often taken to be a *theory* of scientific rationality, and not just a feature of scientific *practice*. As such, it is supposed to provide an account of how researchers should draw conclusions about their hypotheses from data, or which of their epistemic states are justified. This is what's taken to be at stake in the "statistics wars" between the Bayesian and frequentist. The objections to frequentism are objections to p-values as an account of rational posterior credences, or to the conflation of a with statistically significant result with proof that a null hypothesis is false, or to the paradigm's failure to use a researcher's total evidence.

Introductory statistics textbooks and scientific practitioners often distinguish between "descriptive" and "inferential" statistics, and significance tests and p-values are characterized as inferential statistics. The American Psychological Association [2022] defines inferential statistics as "a broad class of statistical techniques that allow inferences about characteristics of a population to be drawn from a sample of data from that population ... these techniques include approaches

for testing hypotheses, estimating the value of parameters, and selecting among a set of competing models." Inferential statistics stand in contrast with "descriptive statistics" which are defined by the APA as "procedures for depicting the main aspects of sample data, without necessarily inferring to a larger population. Descriptive statistics usually include the mean, median, and mode to indicate central tendency, as well as the range and standard deviation that reveal how widely spread the scores are within the sample. Descriptive statistics could also include charts and graphs such as a frequency distribution or histogram, among others."

A similar distinction is made by philosophers, who characterize frequentist methods as "ampliative" (from the Latin *ampliare*: "to enlarge"), because of the potential for frequentist inference to transcend the data, and to "go beyond" or "take the leap from" the data towards justified conclusions [4] In taking this leap, the ampliative methods are fallible — researchers can draw false conclusions by using them. In contrast, the "non-ampliative" methods are infallible: they are used to draw conclusions that follow deductively from the data.

On this understanding of frequentist statistics, the practice of running statistical tests is part of the *context of justification*, the context in which scientific conclusions are drawn from data and observations. This location of frequentist statistics in the context of justification is why significance testing can been used to challenge the value-free ideal. It's the context of justification that raises most divisive questions about the role of non-epistemic values in science. In contrast, *context of discovery*, in which the research subject matter is selected, questions are formed, and data is collected, is unquestionably value-laden. The scientific endeavour to discover a cure for cancer or understand the earth's climate change is certainly motivated by more than intellectual curiosity. The *context of application*, in which scientific findings are communicated, reported, and put into action, is also unquestionably value-laden. Which (and how) scientific discoveries are published, reported in the media, and influence policy are all determined by value-laden choices.

[4] Genin [2020] characterizes this distinction, and Mayo [2011] and Spanos [2010] characterize frequentist inference as ampliative.

The claim that science is "value-free" and isolated from the biases, idiosyncrasies, social locations, and subjectivity of the scientists could only ever be plausible as a claim about the context of justification.[5] If statistical significance testing is a practice in the context of justification, and the selection of significance levels depends on non-epistemic values, then the activity of the context of justification is value-laden.

But this is all to task frequentist statistical methods with more than they could possibly do. For reasons we have already seen, frequentism does not provide a satisfactory account of scientific rationality. P-values are not rational posterior credences in a null hypothesis, and a statistically significant result is not proof that a null hypothesis is false. Bayesianism is better suited as a theory of scientific rationality. Nevertheless, the calls to "abandon statistical significance" [McShane et. al, 2016] or to "ban the p-value" [Kraemer, 2019] equivocate between the following questions:

1. What is the correct theory of scientific rationality?

2. What statistics should we compute in practice?

Suppose we accept Bayesianism as the answer to the first question. The Bayesian scientist has a problem: they cannot just conditionalize on the entire dataset. Scientists face limitations: cognitive, temporal, and spatial. If a researcher skims a large dataset in an attempt to conditionalize on it, it's unlikely that they'll do a very good job. Their eyes may glaze over it and fixate on some data points, and then they learn from a subset of the data. If the data is a large sequence of H's and T's, they may try to count the number of each. If the data is a list of home prices in the US, they may gain a general sense of the numbers: many entries are in the $400,000s. If the data is a spreadsheet of demographic data, they may look for correlations: countries with low GNP seem to have high maternal mortality rates. The researchers will learn more if they calculate

[5] Reichenbach originally drew a distinction between the context of discovery (the context in which hypotheses are formed) and the context of justification (the context in which hypotheses are tested), and the taxonomy of scientific practice that includes the context of discovery, justification, and application follows Anderson [1995].

statistics, for example, the number of Hs and Ts, the median home prices, or a correlation coefficient for GNP and maternal mortality rates. The descriptive statistics are valuable for learning about a dataset, and then drawing conclusions from the data. Frequentist statistics can also be used to help researchers learn about the data. Significance tests condense information in ways that researchers can learn from and that Bayesians can conditionalize on.

The distinction between descriptive and inferential statistics, and between ampliative and non-ampliative statistics, pigeonholes frequentist statistics into a role it cannot perform. The results of a statistical test does not contain any information that is not already in the dataset, or allow researchers to "go beyond" or "take the leap from" the data. Like descriptive statistics, the significance test follows by mathematical deduction, from the data. Significance tests answer a yes/no question about the data: was the observed outcome in the rejection region or not? It's a reduction of the data, and it this the reduction that is useful to researchers, given their bounded rationality. A Bayesian researcher cannot conditionalize on an entire dataset, but they can conditionalize on statistics of the data, including the results of significance tests.

Significance tests answer a yes/no question about the data and the experimental results: was the observed outcome in the rejection region or not? Significance testing is then a practice in the context of discovery and in the context of justification: it's a way of exploring, understanding, and communicating the experimental data. It's how scientists learn about the data they have collected — how they summarize, reduce, and selectively attend to it in a way that helps them learn from it. The result of a significance test is evidence that Bayesian researchers can conditionalize on.

4.2.3 Accuracy and Statistical Significance

There are better and worse ways to summarize information. In the case of our coin flip experiment, it is useful to calculate the number of Heads. And there are less useful statistics to calculate: whether the number of Heads is even or odd, how the coin landed in just the first three flips, or the number of times that the coin landed Heads three times in a row. There are formal properties

of the number of Heads that explain why it's useful, and a better way to summarize information than these other statistics. It's a *sufficient statistic*, which means that the statistic contains everything that is evidentially relevant to the hypothesis. In the Bayesian framework, $T(X)$ is a sufficient statistic if and only if:

$$c(\theta|X) = c(\theta|T(X))$$

All of the information about $X = X_1, \ldots, X_n$ that is evidentially relevant to the value of θ is given by $T(X)$. If a Bayesian updated their credences in the bias of a coin by conditionalizing on the exact sequence of the coin flips, or the by conditionalizing on the number of Heads and Tails, they would arrive at the same posterior credence distributions.

When the p-value is calculated using a sufficient statistic, it may be a sufficient statistic itself. For example, in the coin flip case, if we knew the number of times the coin landed Heads, we could compute the p-value. And if we knew the p-value, we could figure out the number of times the coin landed Heads. The two statistics are equivalent. If you learn the p-value, the number of Heads, or the exact sequence of coin flips, you should end up the same posterior credences in the bias of the coin. In this case, to learn that the experimental results are significant at the .05 level just is to learn that the coin landed Heads at least 20 times.

The significance test reduces the data to one-bit. Formally, you could think of this as a function $t : X \rightarrow \{0, 1\}$, a significance test takes the data, and assigns it a value of 0 if the result is not statistically significant (i.e., the data falls out of the rejection region, or the p-value is less than the α level) or 1 if the result is statistically significant (i.e., the data falls within the rejection region, or the p-value is greater than the α level). Although there is only so much information that can be communicated in one-bit (it's *one bit*, after all) there is some value to reducing data to one-bit. It's easier to remember and communicate whether or not a result is statistically significant than it is to remember or communicate an exact p-value. There is also some value in having a significance level that is set by convention, having one significance level that is used

for all of the studies in a discipline or for some application. Again, it's easier to remember and report, but the standardization is also convenient: if you look at citation practices in psychology textbooks and journals, authors tend to cite lists of experiments that support similar hypotheses with statistically significant results.

If you could ask any one-bit question about a dataset (or have a convention for asking this question about any dataset), what should it be? It depends on your goals — what you want to learn about the data or want you want to learn about your hypotheses. But suppose you value the overall accuracy of the credences you would have in your hypotheses after you learn the answer to the question: you value having a high credence in the true hypothesis, and a low credence in the false hypotheses. Using the Brier score, the inaccuracy of a credence function c for a proposition p can be expressed as $(c(p) - 1)^2$ if p is true and $(c(p))^2$ if p is false.[6] The expected Brier inaccuracy in the value of θ when conditionalizing on the answer to a question t can be given as follows[7]:

$$I(t) = \int_\theta c(\theta)\Big[P_\theta(t) \cdot \big(c(\theta|t) - 1\big)^2 + P_\theta(\bar{t}) \cdot \big(c(\theta|\bar{t}) - 1\big)^2\Big]$$

If we seek to minimize $I(t)$, it turns out that updating on the results of a significance test performs quite well. In fact, in some typical cases, one of the best yes/no questions you could ask about the data is: is the result significant at the .05 level?

For example, consider our running example, of flipping the coin to learn about its bias, and consider a null hypothesis $H_0 : \theta = .5$ and an alternative hypothesis $H_1 : \theta = .8$. For most prior credences assignments over these hypotheses, when $n = 30$, updating on the significance test at the .05 level has a much lower expected inaccuracy than updating at the Bayes factor test at the .1 level (Figure 1), for when $n = 100$, the Bayes factor test has a lower expected inaccuracy, although

[6]The Brier score is often used in arguments to justify the axioms of Bayesianism, and Perez Carballo [2018] uses the Brier score to measure the value of asking questions.

[7]For this definition, c is a probability density function, but a corresponding definition could be given when c is a probability mass function.

both result in low expected inaccuracy (Figure 2). Figure 1 shows the expected inaccuracy of updating on the significance test at the .05 level, a Bayes factor test at the .1 level, the entire data set, and no information at all, for any prior credence distributions over the hypotheses, for $n = 30$. For most of the priors, updating on the significance test at the .05 level has a lower expected inaccuracy than updating on the Bayes factor test at the .1 level. The expected inaccuracy is so low that updating on the significance test closely approximates the expected inaccuracy of updating on the entire data set. Figure 2 shows the expected inaccuracy of updating on the significance test at the .05 level, a Bayes factor test at the .1 level, the entire data set, and no information at all, for any prior credence distributions over the hypotheses, but this time, with a larger sample size of $n = 100$. With the higher sample size, the Bayes factor test has the advantage over the frequentist test, but updating on a significance test still has low expected inaccuracy.

Suppose we consider a different alternative hypothesis, so that we have a have a null hypothesis $H_0 : \theta = .5$ and an alternative hypothesis $H_1 : \theta = .6$. Updating on the significance test at the .05 level has a lower expected inaccuracy than updating on the Bayes factor test at the .1 level, for any prior credence over the hypotheses, and when $n = 30$ or $n = 100$. Figure 3 shows the expected inaccuracy of updating on the significance test at the .05 level, a Bayes factor test at the .1 level, the entire data set, and no information at all, for any prior credence distributions over the hypotheses, for $n = 30$. For any of the priors, updating on the significance test at the .05 level has a lower expected inaccuracy than updating on the Bayes factor test at the .01 level. Figure 4 shows the expected inaccuracy of updating on the significance test at the .05 level, a Bayes factor test at the .1 level, the entire data set, and no information at all, for any prior credence distributions over the hypotheses, but this time, with a sample size of $n = 100$. In all of these cases, conditionalizing on a statistical significance test at the .05 level results in a lower expected inaccuracy than conditionalizing on a Bayes factor test at the .1 level. So accuracy considerations alone — at least in these cases — are favorable to statistical significance testing. Bayesian updating on the result of a Bayes factor test at the .1 level has less expected accuracy than updating on the significance

test at the .05 level.

This approach also explains a curious fact about the practice of significance testing: the differences in the conventional significance level across academic disciplines. The proponents of the standard approach to setting significance levels often acknowledge that, in practice, significance levels are not set in accordance with the cost of errors: "Under most circumstances, the choice of a level of statistical significance is not made through the explicit consideration of arguments for different statistical choices, but by the tradition of an area of research or the choice of a computer statistical package [Douglas, 2020]." Statisticians often hold that these values have no theoretical justification at all: "In practice three levels are commonly used: 1 percent, 5 percent and 0.3 of one percent. There is nothing sacred about these three values; they have become established in practice without any rigid theoretical justification [Fisher, 1933]." but the expected accuracy of using these significance levels can explain these values.

Significance levels are often set at .05 in the psychological and social sciences, .01 in the biological and medical sciences, and $< .001$ in the physical and computational sciences. Here's the interesting point. As sample sizes increase, the optimal significance level gets lower. Figures 5 - 11 show the significance level that minimizes expected inaccuracy for hypotheses under consideration for each sample size. And we can see that when $n = 30$, which is the typical sample size in the psychological and social sciences, the optimal significance level is approximately .05. When the sample size is 100, which is typical in biological and medical research, the optimal significance level is approximately .01, and in the computational sciences, where sample sizes can be exceedingly large, the significance levels are exceedingly low.

The accuracy considerations could be part of a rationalizing explanation for the practice of computing statistical significance tests. It could be part of the explanation for why the practice has continued, after behaviorism and falsificationism have fallen to the wayside, after significance testing has fallen under widespread criticism, and after Bayesianism has become a widely accepted theory of scientific rationality. Frequentist statistics are useful to the Bayesian. For

the Bayesian to "abandon statistical significance" is to dispense with a potentially useful tool. A Bayesian can learn, rationally, from frequentist statistics.

Of course, there are limits to this rationalizing explanation — to both the mathematical results and to the value of frequentist statistics. First, the mathematical results consider one statistical model. The Bernoulli distribution is just one of many statistical models that are used in practice. For some test statistics, this does not matter much: according to the Central Limit Theorem, the sampling distribution of a sample mean will converge in distribution to the normal distribution. Or, in simplified terms: when $n \geq 30$, the distribution of a sample mean will approximate a normal distribution, regardless of the distribution in the underlying statistical models. Yet there are a wide range of statistical models used in practice. The question of how well significance tests condense information in these cases is a fascinating mathematical question (and beyond the scope of this paper).

Second, the Neyman-Pearson lemma holds for our statistical models, which explains some of the success of the frequentist significance test in the running example — the statistical significance test is equivalent to some Bayes factor test. According to the Neyman-Pearson lemma, in cases in which a uniformly most powerful test exists (as in our example), the significance test will be equivalent to reporting a Bayes factor threshold, at some level. So we could translate a significance test into a Bayes factor threshold, at some level, and vice versa. So the comparison between a Bayes factor threshold and a frequentist significance test, in these cases, is a comparison between a convention of holding fixed a Bayes factor threshold or holding fixed a significance level.

Finally, in actual scientific practice, researchers are unlikely to have all of their credence in two pointwise hypotheses. The example is an oversimplification, but may approximate actual credence distributions: $H_1 : .8$ could stand in for a credence distribution over $[.5, 1]$ that is centered around (or near) .8, and $H_1 : .6$ could stand in for a credence function that is centered around .6. The significance test has uses and limitations. Although significance tests condense

data efficiently, they are probably not the *best* one-bit condensation of the data (they're good, but probably not the *best*.) Furthermore, it may be the case that they are so misunderstood by researchers and that it's so tempting to confuse p-values with posterior credences that it is best to abandon them. Given the nature of human cognition, it may turn out that some other way of condensing data is more useful. And in the end, a significance test reduces the set one-bit of information, and this may just be too condensed. It may be worth the loss of economy to communicate and learn from more than one-bit of information.

Here's the important lesson. The question of how to set significance levels is not moot if we accept a Bayesian theory of scientific rationality. First, it does not follow from a Bayesian theory of scientific rationality that the practice of running and reporting significance tests should end. In fact, the practice could be useful to the Bayesian. Second, the Bayesian researcher is currently in a frequentist world. If significance tests continue to be reported, that's the evidence that the Bayesian will conditionalize on. The Bayesian will learn different things, and come to hold different posterior credence distributions, depending on which significance tests they conditionalize on.

4.3 How To Set Significance Levels

How should the Bayesian set significance levels? In the previous section, we saw that significance levels could be set in a way that allows researchers to minimize expected inaccuracy. So if minimizing expected inaccuracy is the goal, then we've already made progress on the question. However, minimizing expected inaccuracy in a hypothesis may not be the only worthwhile goal for researchers. And it could be perfectly rational to inquire in a way that yields less expected accuracy because another posterior credence distribution would be more useful.

Here's an example. Suppose you're a prosecutor investigating a murder and the courthouse that's housing your evidence is burning into the ground. You only have time to go back in and

save one piece of evidence: one of two videotapes that you haven't watched yet. You have two suspects: Professor Plum and Colonel Mustard. The first videotape captured the room with the murder, and so certainly caught the murderer on tape. But it's grainy: you could figure out from the tape whether the murderer wore a purple or yellow jacket, which is good evidence for who did it, but not strong enough to secure a guilty verdict. The second videotape captured one of the doors into the mansion. If the murderer happened to enter through that door, they would be caught on tape, and suppose it's a clear tape, so you would certainly know who did it, and could secure a guilty verdict. Which tape should you save? If your primary aim is to have evidence that supports a high credence — or knowledge — in the suspect's guilt, you should save the clear tape of one of the entrances to the mansion. It may give you less expected accuracy than if you saved the grainy tape, but it'll maximize the chance that you get what you care about, which is to secure a guilty verdict.

So here's what I think accounts for the relationship between values (both epistemic and non-epistemic) and significance levels:

1. How we should set significance levels depends on the value of the posterior credence functions that could result from conditionalization on the significance test.

2. The value of a posterior credence function depends on epistemic and non-epistemic values.

Statistical significance testing is a way of asking questions about a data set. The questions scientists should ask depend on the value of the answers.

How to set significance levels? This section will work through two examples. First, the paradigm case of values in significance testing: how to set significance levels in testing drugs for safety and efficacy. Second: the case that motivated Neyman-Pearson significance testing: the search for differences between populations, originally in scientific eugenics programs, and later in "scientific racist" and "neurosexist" research programs. In the first example, the Stan-

dard Account makes the correct recommendations. In the second example, the Standard Account does not make the correct recommendations, and instead makes recommendations that may be seriously problematic.

For both examples, I'll continue to use the running example of our Bernoulli statistical model (n=30), with $H_0 : .5$ ad $H_1 : .8$, and $c(H_0) = c(H_1) = .5$.

The table shows the possible posterior credences in H_0 that result from conditionalizing on a significance test at the .01, .05, or .5 levels.

α	reject	fail to reject
.01	<.01	0.86
.05	.048	.97
.5	.36	>.99

4.3.1 Example 1: FDA approval and drug testing

4.3.1.1 Proof of Safety

The FDA requires pharmaceutical manufacturers to provide "proof" of a drug's safety prior to approval.[8] The drug is often tested on animals for toxicity, before it moves to Phase I clinical trials, where the drug is tested for toxicity on approximately 20-80 subjects. Suppose the null hypothesis is that the drug is safe — the treatment group and the control group have the same rates of adverse medical outcomes. How should the regulators set statistical significance levels?

According to the Standard Account, the researchers could make two errors. A Type I error: they reject a true null hypothesis, and do not approve a safe drug (overregulation). A Type II error: they fail to reject a false null hypothesis, and approve an unsafe drug (underregulation). Each of these errors has a cost. If the safe drug is not approved, then patients could miss out on a effective treatment. If the unsafe drug is approved, then patients could experience adverse

[8] See Mantus & Pisano [2017] for a comprehensive overview of the FDA drug approval process.

medical outcomes. Let's say Type II error is the more costly error. Then according to the Standard Account, significance levels should be set high.

How do I think we should set significance levels? Suppose that it's much more costly to approve an unsafe drug than it is to not approve a safe drug. Then, the regulators need to be confident that it's a safe drug in order to approve it. If we are standard decision theorists, the credence required in order to approve the drug will depend on the utilities of the outcomes. So let's say that the utilities are given as follows:

	safe	not safe
approve	10	-100
do not approve	0	0

Then, in order for the expected utility of approving the drug to exceed the expected utility of not approving the drug, the regulators must have at least a .91 posterior credence that the drug is safe. Suppose their posterior credence distribution align with our running example. If they could ask one yes/no question about the data, it should be: does the data support at least a .91 posterior credence that the drug is safe? So they should set their significance levels somewhere between .01 and .05.

Suppose we raise the cost of approving an unsafe drug, so that the utilities of the outcomes is given as follows:

	safe	not safe
approve	10	-1000
do not approve	0	0

This time, in order for the expected utility of approving the drug to exceed the expected utility of not approving the drug, the regulators must have at least a .99 posterior credence that the drug is safe. If they could ask one yes/no question about the data, it should be: does the data support a .99 credence that the drug is safe? So they should set the significance levels higher and close to

.5. In this case, the result of a significance test at a lower level is not helpful for their deliberation. If the regulators are only told whether or not the result is significant at the .01 or .05 level, they should not approve the drug in either case: their credence that the drug is safe will not reach .99.

This all accords with the Standard Account. As the Type II error becomes more costly, the significance levels should be raised. This all does not, however, accord with the current FDA practice, in which significance levels are set at the .05 or .01 level. The regulators can end up with a high posterior credence that a drug is *unsafe* (i.e., proof of hazard), but not with a high posterior credence that a drug is *safe*.[9] This should raise eyebrows about the current FDA practice, but so far, the Standard Account is doing just fine.

4.3.1.2 Substantial Evidence of Efficacy

The FDA also requires pharmaceutical manufacturers to provide "substantial evidence" of a drug's efficacy. The FDA Guidance For Researchers states: "The strength of evidence in each trial contributing to meeting the substantial evidence standard should be assessed by appropriate statistical methods. The uncertainty about the findings from each trial should be sufficiently small and the findings should be unlikely to result from chance alone, as demonstrated by a statistically significant result or a high posterior probability of effectiveness." Although the FDA typically requires at least two "adequate and well-controlled" trials, as a single study can also suffice if it of sufficient quality or accompanied by "convincing" evidence of the drug's mechanism of action in treating a disease or condition.

How would the Standard Account treat this case? The null hypothesis is that the drug is not effective — the treatment and control groups have outcomes. The researchers could make two errors. A Type I error: they reject a true null hypothesis, and they approve a drug that is not effective. A Type II error: they fail to reject a false null hypothesis, and they do not approve a

[9] See Liu [2017] for a discussion of the "proof of safety" and "proof of hazard" standards with an argument for why we should favor the former and Stegenga [2018] for an argument that medical research underestimates the risk of the harms of therapeutics.

drug that is effective. Each of these errors has a cost. If the ineffective drug is approved, patients endure side-effects and other costs in order to take a drug that doesn't even work. If the effective drug is not approved, patients miss out on an effective treatment (or, more realistically, the pharmaceutical manufacturers need to gather more evidence in order to demonstrate the efficacy of the drug.) Let's say the Type I error is more costly, so according to the Standard Account, the significance levels should be set at a low level.

How do I think we should set significance levels? Suppose that it's more costly to approve an ineffective drug than it is to not approve an effective drug. Then, the regulators need to be confident that it's effective in order to approve it. Again, if we are standard decision theorists, exactly how confident will depend on the utilities of the outcomes. Let's say that the utilities are as follows:

	effective	not effective
approve	10	-50
do not approve	0	0

And let's continue with our running example of a statistical model. The simplicity of the model, and the $n = 30$ sample size are now quite unrealistic, so this is now a general methodological point. The model could be swapped out to one that's more realistic, but the method will remain the same. In order for the approval of the drug to have the highest expected utility, the regulators must have an at least .83 credence that the drug is effective, so an at most .17 credence in the null hypothesis that the drug is not effective. So they should set significance levels somewhere between .05 and .5.

Now suppose that the Type II error is more costly,

	effective	not effective
approve	10	-50
do not approve	0	0

This time, in order for the approval of the drug to have the highest expected utility, the regulators must have an at least .98 credence that the drug is effective, so an at most .02 credence in the null hypothesis that the drug is not effective. So they should set significance levels lower, somewhere between .01 and .048.

So my account of how to set significance levels and the Standard Account agree in this case. The more costly it is to approve an ineffective drug, the lower the significance levels should be.

4.3.2 Example 2: Scientific Racism and Neurosexism

Fisher's exact test has a charming history: it was motivated by a friendly visit with his friend, Lady Bristol and her talent for tasting tea. Neyman-Pearson significance testing has a dark history. Egon Pearson held the Galton Chair of Eugenics at the University College of London, and was the editor of the Annals of Eugenics. He developed his statistical methods to protest Jewish immigration into Britain. His research program had the aim of discovering difference between Jewish immigrants and the rest of the population and he wrote: "taken on the average, and regarding both sexes, this alien Jewish population is somewhat inferior physically and mentally to the native population [Pearson & Moul, 1925]" The guiding values of this project are clear and explicit. Galton, who coined the term "eugenics" defined the academic discipline as: "the study of agencies under social control that may improve or impair the racial qualities of future generations either physically or mentally." Race science continues in the contemporary US context, and scientists search for racial differences in traits like IQ, using the statistical methods developed by Fisher, Neyman, and Pearson.

In an active body of research wryly called "neurosexist," psychologists are on the search for sex differences in the brain, to explain pervasive and manifest gender differences in our social world. In the Victorian era, scientists measured and male and female brains, and concluded that men had the advantage of an extra "five ounces" of grey matter. Now, this project is fueled by fMRI research, which allows for more sophisticated hypotheses about male and female brains.

For example, men's brains are hypothesized to feature more lateral connectivity in the right side of the brain, which is supposed to explain men's superiority at spatial reasoning, while women's brains are hypothesized to feature more lateral connectivity in the left side of the brain, and these differences are then used to explain women's adept language skills [Tomasi & Volkow, 2011]. These differences are then further used to explain — or justify — gender roles and patterns in occupations, political representation, domestic and reproductive labor, and violence [Fine, 2010]. Some feminist scholars have suggested that the danger of accepting these hypotheses could justify raising the evidential standards for accepting these hypotheses [Hare & Mustin, 1994].

The moral landscape is difficult terrain. Surely, even if the differences in IQ (or hygiene or dress or facial structure or any other features that Pearson was researching) exist, they would not justify a discriminatory and exclusionary immigration policy, or the atrocities that this research fueled. The fMRI scans of male and female brains can uncover sex dimorphisms, but conclusions about the causes of the differences cannot be made on the basis of the scans alone. It's one thing to find dimorphisms, but it's quite a leap to conclude that they are an inevitable expression of chromosomes, and not a result of the interaction with the social environment. And it's an even further leap to think that any of these studies can justify the patriarchal *status quo*. In addition, there can be value in finding differences, where they exist, for the project of building a more equitable social world.

How would the Standard Account treat these cases? Suppose the Type I error (to conclude that there are differences, when there are none) is the more costly mistake. And suppose that the Type II (to conclude that there are no differences, when there are differences) error is less costly. If this is the case, then according to the Standard Account, we should lower the significance levels, to lower the Type I error rate and increase the Type II error rate.

But there's a moral complexity that the Standard Account cannot accommodate. Suppose you think that if there is a research program in racial or sex differences, it should be possible to come to know, on the basis of this research, that the differences do not exist (if in fact they do not.) Maybe

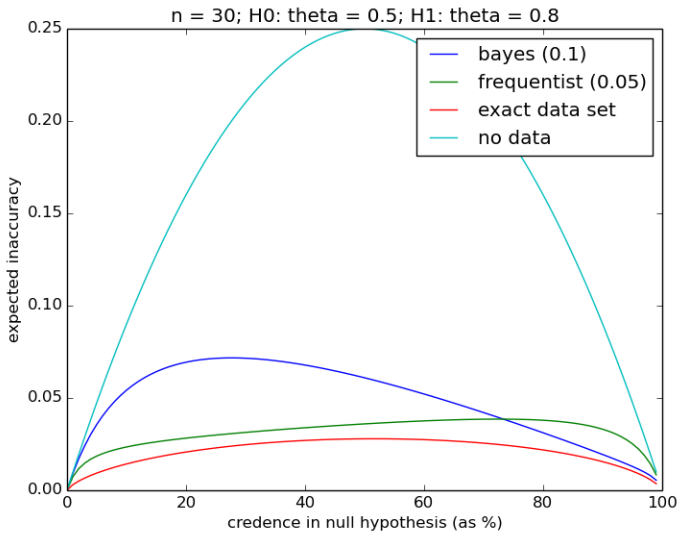
you think that it's better to have a research program in which we could prove there are similarities than a research program in which we could prove that there are differences, but never prove that there are similarities. If this is the case, then we should *raise* the significance levels, and not *lower* them. Consider our running example. If we *lower* the significance levels (say, to .01), then rejecting the null hypothesis will result in an extremely high posterior credence in the alternative hypothesis (.99), and failing to reject the null hypothesis will lead to a moderately high credence (.86) in the null hypothesis. So if the credal threshold for knowledge is approximately .95, then it is possible to know that the null hypothesis is false, but not that it is true. The significance test at the .01 level can produce strong evidence against the null hypothesis, but not strong evidence for the null hypothesis.

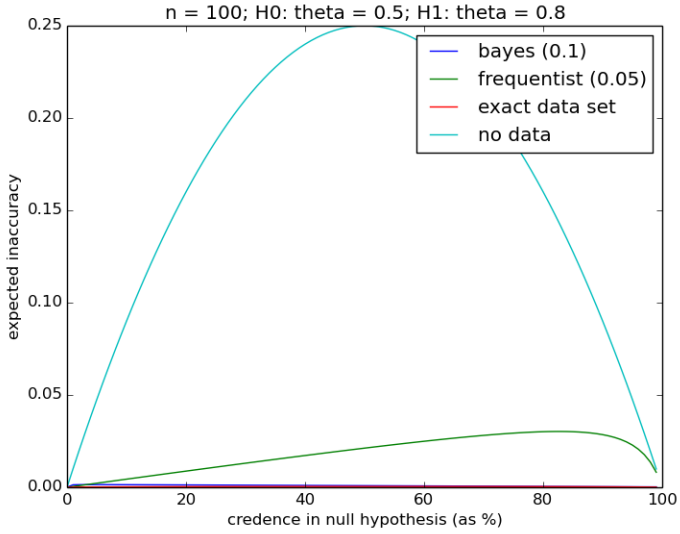
Pearson developed significance testing (and the Standard Account) in order to discover differences between populations. This is exactly what the tools can do. They can be used to discover differences, but not similarities. The Standard Account of how to set significance levels only exacerbates the problem. The Standard Account makes us less likely to conclude that there are differences, but with even more certainty when when we do.

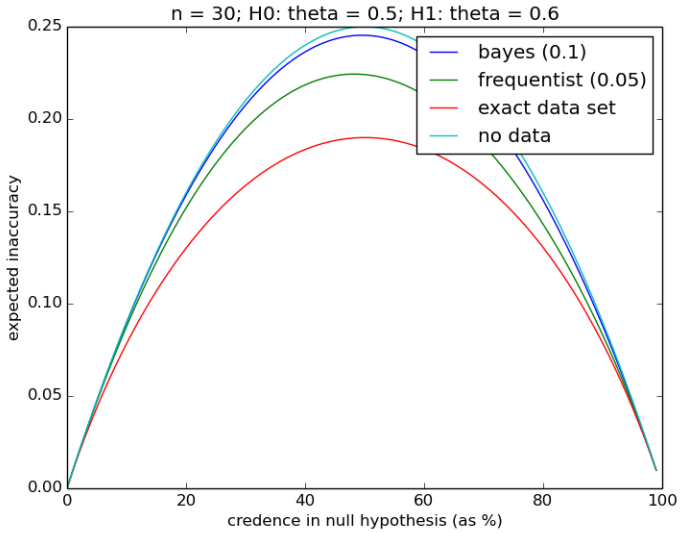
4.4 Conclusion

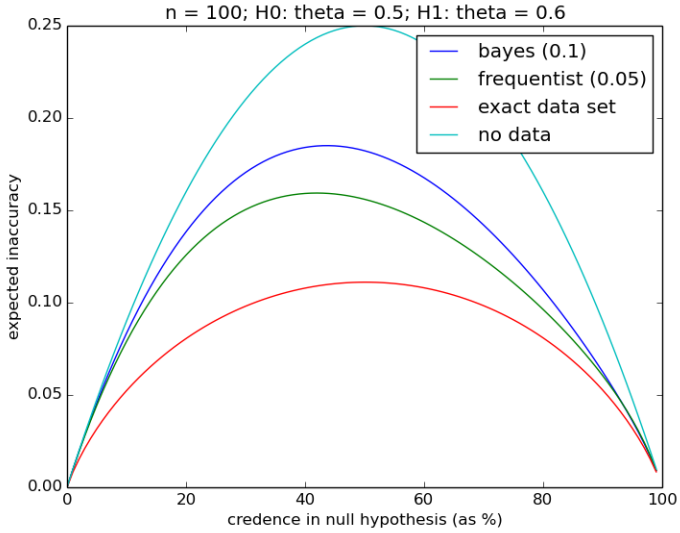
How should we set significance levels? Here's the short answer. To run a significance test is to as one yes/no question about your dataset, which you can then conditionalize on. The significance levels make a difference to which posterior credences you may end up with. The value of a posterior credence function depends on features deliberative context — your options and their practical stakes. So look at your options, figure out which posterior credences distributions can help you decide what to do, and then figure out which statistical levels can get you those credences.

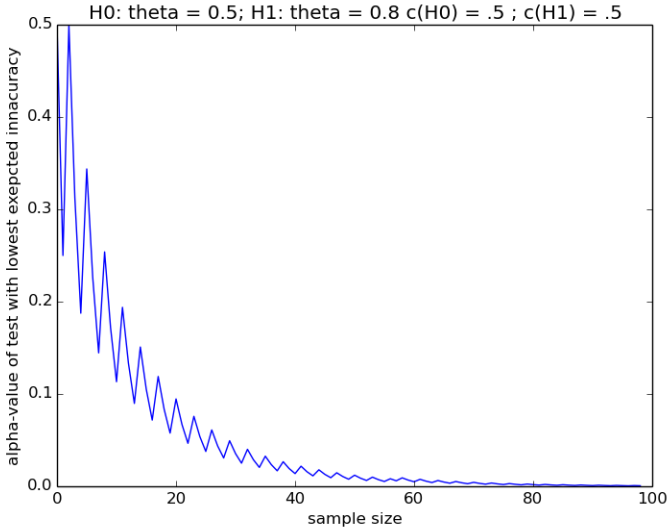
4.5 Figures

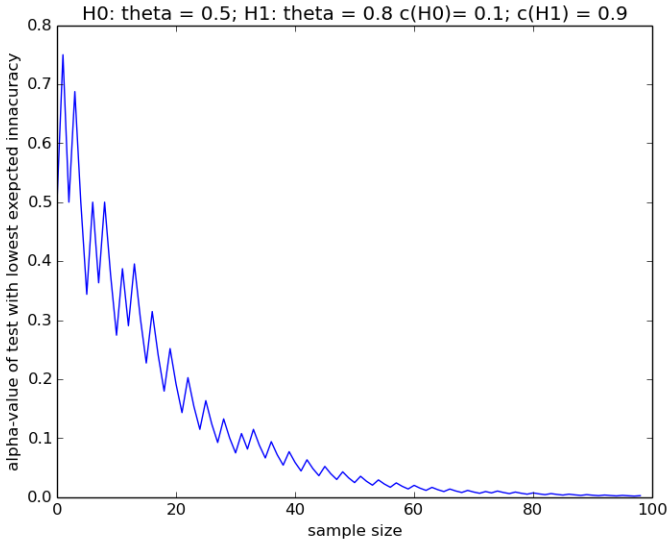

n = 30; H0: theta = 0.5; H1: theta = 0.8
bayes (0.1)
frequentist (0.05)
exact data set
no data
expected inaccuracy
credence in null hypothesis (as %)
0.00
0.05
0.10
0.15
0.20
0.25
0
20
40
60
80
100

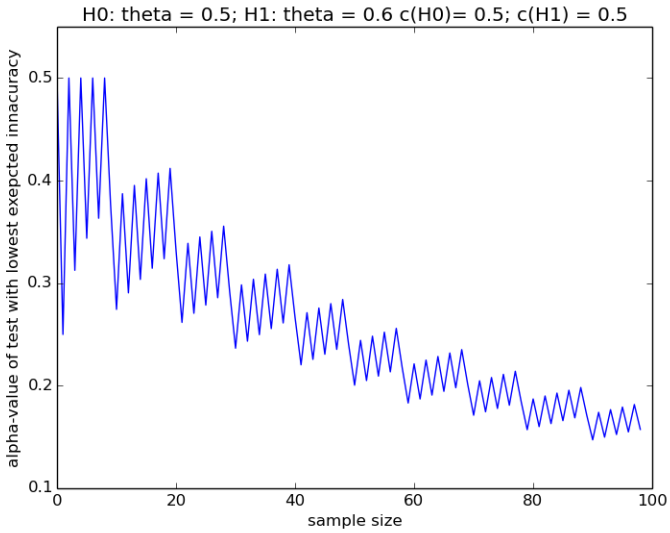

n = 100; H0: theta = 0.5; H1: theta = 0.8
bayes (0.1)
frequentist (0.05)
exact data set
no data
expected inaccuracy
credence in null hypothesis (as %)
0.00
0.05
0.10
0.15
0.20
0.25
0
20
40
60
80
100

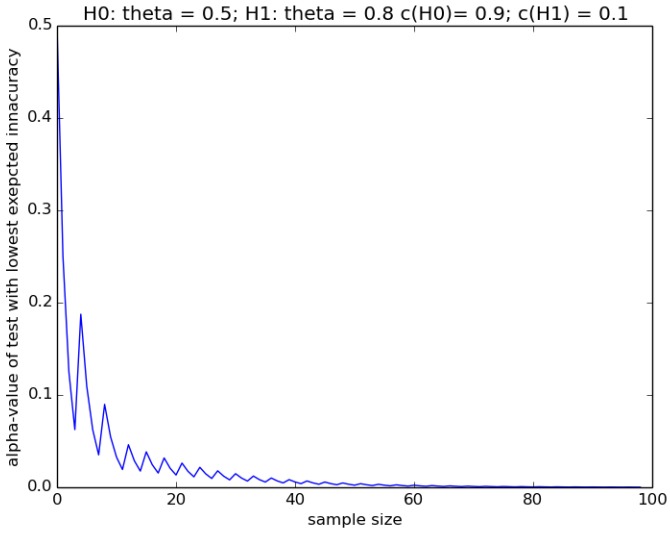

n = 30; H0: theta = 0.5; H1: theta = 0.6
bayes (0.1)
frequentist (0.05)
exact data set
no data
expected inaccuracy
credence in null hypothesis (as %)
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100

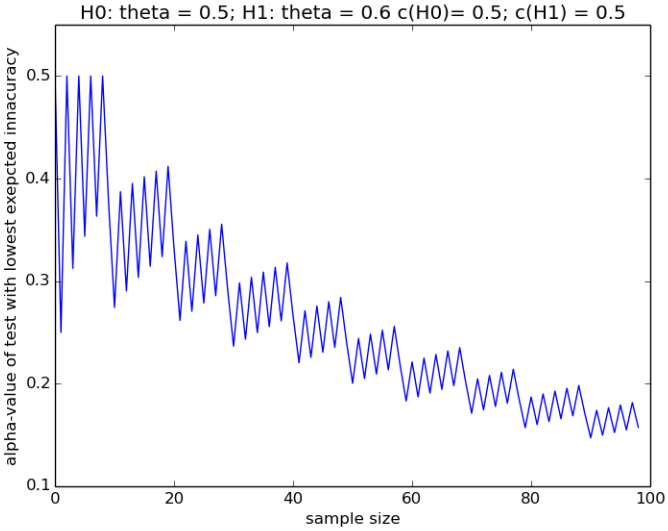

n = 100; H0: theta = 0.5; H1: theta = 0.6
bayes (0.1)
frequentist (0.05)
exact data set
no data
expected inaccuracy
credence in null hypothesis (as %)
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100

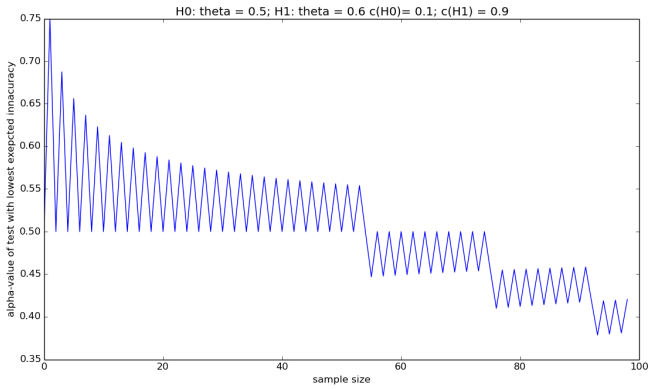

H0: theta = 0.5; H1: theta = 0.8 c(H0) = .5 ; c(H1) = .5
alpha-value of test with lowest exepcted innacuracy
sample size
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
100

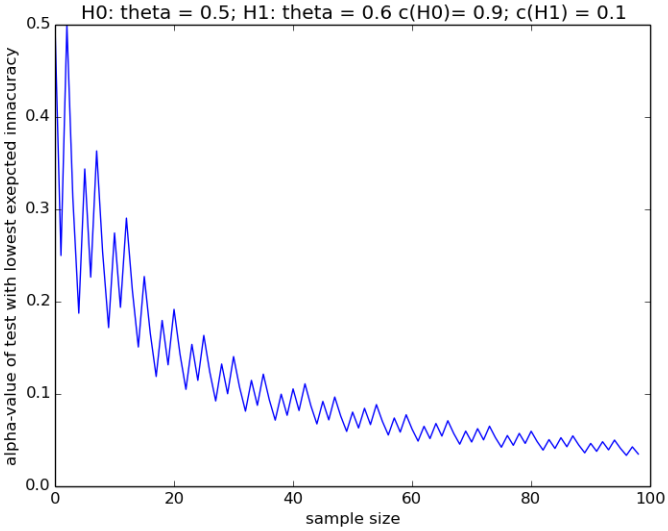

H0: theta = 0.5; H1: theta = 0.8 c(H0)= 0.1; c(H1) = 0.9
alpha-value of test with lowest exepcted innacuracy
sample size
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
20
40
60
80
100


H0: theta = 0.5; H1: theta = 0.6 c(H0)= 0.5; c(H1) = 0.5
alpha-value of test with lowest exepcted innacuracy
sample size
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
100


H0: theta = 0.5; H1: theta = 0.8 c(H0)= 0.9; c(H1) = 0.1
alpha-value of test with lowest exepcted innacuracy
sample size
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
100


H0: theta = 0.5; H1: theta = 0.6 c(H0)= 0.5; c(H1) = 0.5
alpha-value of test with lowest exepcted innacuracy
sample size
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
100


H0: theta = 0.5; H1: theta = 0.6 c(H0)= 0.1; c(H1) = 0.9
alpha-value of test with lowest exepcted innacuracy
sample size
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0
20
40
60
80
100


H0: theta = 0.5; H1: theta = 0.6 c(H0)= 0.9; c(H1) = 0.1
alpha-value of test with lowest exepcted innacuracy
sample size
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
100

Citations

Aler, Tubella, Andreas Theodorou, Virginia Dignum, and Loizos Michael. "Contestable Black Boxes." *In Rules and Reasoning*, edited by Víctor Gutiérrez-Basulto, Tomáš Kliegr, Ahmet Soylu, Martin Giese, and Dumitru Roman, 159–67. Springer, 2020.

Anderson, Elizabeth. "Knowledge, Human Interests, and Objectivity in Feminist Epistemology." Philosophical Topics 23, no. 2 (1995): 27–58.

Aneesh, A *Virtual Migration: The Programming of Globalization.* Duke University Press, 2006.

"APA Dictionary of Psychology." Accessed August 6, 2022. https://dictionary.apa.org/. Banerjee, Amitav, Ub Chitnis, Sl Jadhav, Js Bhawalkar, and S Chaudhury. "Hypothesis Testing, Type I and Type II Errors." *Industrial Psychiatry Journal* 18, no. 2: 127, 2009.

Attali, Yigal and Jill Burstein. "Automated Essay Scoring With E-Rater® V.2." The Journal of Technology, Learning and Assessment 4, no. 3, 2006.

Barocas, Solon and Andrew Selbst, "Big Data's Disparate Impact." California Law Review, 2016.

Bell, Kristen. "Toward a Normative Theory of Parole Grounded in Agency." *Philosophical Issues* 31, no. 1: 24–40, 2021

Blackstone, William. *Commentaries on the Laws of England: In Four Books.* Callaghan, 1884.

Bloome-Tillmann, Michael. *Knowledge and Presuppositions.* Oxford University Press, 2014.

——— “More Likely Than Not" - "Knowledge First and the Role of Statistical Evidence in Courts of Law.” *In Knowledge First: Approaches in Epistemology and Mind*, edited by J. Adam Carter, Emma C. Gordon, and Benjamin W. Jarvis, 278–92. Oxford: Oxford University Press, 2017.

Brown, George W. “Errors, Types I and II.” Archives of Pediatrics Adolescent Medicine 137, no. 6: 586, 1983.

Brown, Jessica. “Subject-Sensitive Invariantism and the Knowledge Norm for Practical Reasoning.” *Nous* 42, no. 2: 167–89, 2008

Buchak, Lara. “Belief, Credence, and Norms.” *Philosophical Studies* 169, no. 2: 285–311, 2014

Busuioc, Madalina. “Accountable Artificial Intelligence: Holding Algorithms to Account.” *Public Administration Review* 81, no. 5: 825–36, 2021

Chan, Stephanie, Vidhatha Reddy, Bridget Myers, Quinn Thibodeaux, Nicholas Brownstone, and Wilson Liao. “Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations.” *Dermatology and Therapy* 10, no. 3: 365–86, 2020

Coeckelbergh, Mark. *The Political Philosophy of AI: An Introduction.* Polity, 2022.

Cohen, Stewart. “Contextualism, Skepticism, and the Structure of Reasons.” *Nous* 33, no. s13: 57–89, 1999

———. “How to Be a Fallibilist.” *Philosophical Perspectives* 2: 91, 1999.

Colaner, Nathan. “Is Explainable Artificial Intelligence Intrinsically Valuable?” *AI & Society* 37, no. 1: 231–38, 2022

———. “Is Explainable Artificial Intelligence Intrinsically Valuable?” *AI & Society* 37, no. 1: 231–38, 2022.

Corbett-Davies, Sam, and Sharad Goel. “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.” arXiv, August 14, 2018.

Coughlan, Sean, *BBC News*, "Why Did the A-Level Algorithm Say No?," August 14, 2020.

Cranor, Carl F. Regulating Toxic Substances: A Philosophy of Science and the Law. New York, N.Y.; Oxford: Oxford University Press, 1997.

Danaher, John. "Freedom in an Age of Algocracy." *In The Oxford Handbook of Philosophy of Technology*249–72. edited by Shannon Vallor. Oxford University Press, 2022.

———. "The Threat of Algocracy: Reality, Resistance and Accommodation." *Philosophy & Technology* 29, no. 3: 245–68, 2016.

DeRose, Keith. "Contextualism and Knowledge Attributions." *Philosophy and Phenomenological Research* 52, no. 4: 913, 1992.

——— *The Case for Contextualism: Knowledge, Skepticism, and Context.* Oxford: Clarendon press, 2009.

Doshi-Velez, Finale and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning." arXiv, March 2, 2017.

Douglas, Heather. "Inductive Risk and Values in Science." Philosophy of Science 67, no. 4 (2000): 559–79.

Dymitruk, Maria. "The Right to a Fair Trial in Automated Civil Proceedings." *Masaryk University Journal of Law and Technology* 13, no. 1: 27–44, 2019.

Elewski, B., S. Brand, T. Degenhardt, S. Curelop, R. Pollak, R. Schotzinger, A. Tavakkol, et al. "A Phase II, Randomized, Double-blind, Placebo-controlled, Dose-ranging Study to Evaluate the Efficacy and Safety of VT-1161 Oral Tablets in the Treatment of Patients with Distal and Lateral Subungual Onychomycosis of the Toenail*." *British Journal of Dermatology* 184, no. 2 270–80, 2021.

Enoch, David, Levi Spectre, and Talia Fisher. "Statistical Evidence, Sensitivity, and the Legal Value of Knowledge." *Philosophy & Public Affairs* 40, no. 3: 197–224, 2012.

Fantl, Jeremy, and Matthew McGrath. "Evidence, Pragmatics, and Justification." *The Philosophical Review* 111, no. 1: 67, 2002.

———. *Knowledge in an Uncertain World*, Oxford Universiry Press, 2009.

Feinberg, Joel. "Justice and Personal Desert." In *Rights and Reason: Essays in Honor of Carl Wellman, edited by Marilyn Friedman, Larry May, Kate Parsons, and Jennifer Stiff*, 221–50. Springer, 2000.

Fine, Cordelia. Delusions of Gender: How Our Minds, Society, and Neurosexism Create Difference. 1st ed. New York: W. W. Norton, 2010.

Gade, Krishna, Sahin Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. "Explainable AI in Industry: Practical Challenges and Lessons Learned." In *Companion Proceedings of the Web Conference*, 303–4. Association for Computing Machinery, 2020.

Gardiner, Georgi. "The Reasonable and the Relevant: Legal Standards of Proof." *Philosophy Public Affairs*, 47, no. 3: 288–318, 2019

Goldman, Alvin "A Causal Theory of Knowing." *The Journal of Philosophy* 64, no. 12: 357. 1967

Gordon, Alexander, Linlin Chen, Galina Glazko, and Andrei Yakovlev. "Balancing Type One and Two Errors in Multiple Testing for Differential Expression of Genes." *Computational Statistics & Data Analysis* 53, no. 5: 1622–29, 2009.

Gordon, Shira E. "Solitary Confinement, Public Safety, and Recidivism Note." *University of Michigan Journal of Law Reform* 47, no. 2 495–528, 2013

Gunning, David, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. "DARPA's Explainable AI (XAI) Program: A Retrospective." *Applied AI Letters* 2, no. 4: e61, 2021

Hamada, Chikuma. "Statistical Analysis for Toxicity Studies." *Journal of Toxicologic Pathology* 31, no. 1: 15–22, 2018.

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

Hawthorne, John, Jason Stanley, and Journal of Philosophy, Inc. "Knowledge and Action:" *Journal of Philosophy* 105, no. 10 (2008): 571–90.

Herrmann, Anne, ed. "Gender And The Meaning Of Difference: Postmodernism And Psychology." *In Theorizing Feminism: Parallel Trends in the Humanities and Social Sciences*, 2. ed., 49–74. Westview, 2001.

Holt v. United States, 218 U.S. 245 (1910)

Holzinger, Andreas. "From Machine Learning to Explainable AI." In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 55–66, 2018.

Holland, Burt, and Nasser K. Ordoukhani. "Balancing Type I and Type II Error Probabilities: Further Comments on Proof of Safety vs Proof of Hazard." *Communications in Statistics - Theory and Methods*, 3557–70, 1990.

Ifenthaler, Dirk. "Automated Essay Scoring Systems." In *Handbook of Open, Distance and Digital Education*, 1–15. Springer, 2022.

Johnson, Deborah G., and Priscilla M. Regan, eds. *Transparency and Surveillance as Sociotechnical Accountability: A House of Mirrors.* Routledge, 2014.

Kamath, Uday, and John Liu. *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*, Springer 2021.

Kaminski, Margot E. "The Right to Explanation, Explained," *Berkeley Technology Law Journal*, Vol. 34, No. 1, 2019.

Kaminski, Margot E., and Jennifer M. Urban. "The Right to Contest AI" *Columbia Law Review* 121, no. 7 (2021): 1957–2048.

Katsh, M. Ethan, and Orna Rabinovich-Einy. *Digital Justice: Technology and the Internet of Disputes.* Oxford University Press, 2017.

Kim, Brian. "In Defense of Subject-Sensitive Invariantism." *Episteme* 13, no. 2, 233–51, 2016

Kraemer, Helena Chmura. "Is It Time to Ban the P Value?" *JAMA Psychiatry* 76, no. 12: 1219–20, 2019

Lamarque, Peter, and Peter Unger. "Ignorance: A Case for Scepticism." *The Philosophical Quarterly* 26, no. 105: 369, 1976

Latessa, Edward J., Shelley L. Johnson, and Deborah Koetzle. *What Works (and Doesn't) in Reducing Recidivism.* Routledge, 2020.

Laudan, Larry. "The Presumption of Innocence: Material or Probationary??" *Legal Theory*: 333–61, 2005.

Lewis, David. "Elusive Knowledge." *Australasian Journal of Philosophy* 74, no. 4: 549–67, 1996.

Lewis, David K. *Counterfactuals.* Blackwell Publishers, 2001.

Lieberman, Matthew D., and William A. Cunningham. "Type I and Type II Error Concerns in FMRI Research: Re-Balancing the Scale." *Social Cognitive and Affective Neuroscience* 4, no. 4: 423–28, 2009.

Littlejohn, Clayton. "Truth, Knowledge, and the Standard of Proof in Criminal Law." *Synthese* 197, no. 12 5253–86, 2020.

Liu, Jen-pei. "Rethinking Statistical Approaches to Evaluating Drug Safety." *Yonsei Medical Journal* 48, no. 6 895, 2007.

London, Alex John. "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability." *Hastings Center Report* 49, no. 1: 15–21, 2019

Lorenz, Lukas, Albert Meijer, and Tino Schuppan. "The Algocracy as a New Ideal Type for Government Organizations: Predictive Policing in Berlin as an Empirical Case." Information Polity 26, no. 1: 71–86, 2021.

Mantus, David, and Douglas J. Pisano, eds. *FDA Regulatory Affairs.* Third edition. CRC Press, Taylor Francis Group, 2014.

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. "Abandon Statistical Significance." *The American Statistician* 73, no.1 235–4, 2019.

Moran, Frederick A. "The Origins of Parole Section II: Origins of Social Thinking in Crime Treatment." *Yearbook 1945*: 71–98, 1945.

Moskvitch, Katia. "Penal Code: The Coming World of Trial by Algorithm." *New Scientist* 219, no. 2933: 36–39, 2013.

Moss, Sarah. "IX—Moral Encroachment." *Proceedings of the Aristotelian Society* 118, no. 2: 177–205, 2018.

———. "Pragmatic Encroachment and Legal Proof." *Philosophical Issues* 31, no. 1: 258–79, 2021.

———. *Probabilistic Knowledge*. Oxford University Press, 2018.

Müller, Vincent C. "Ethics of Artificial Intelligence and Robotics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, 2021.

Naismith, Ben, Na-Rae Han, and Alan Juffs. "Accurate Measurement of Lexical Sophistication with Reference to ESL Learner Data," n.d., 7.

Nesson, Charles R. "Reasonable Doubt and Permissive Inferences: The Value of Complexity." *Harvard Law Review* 92, no. 6: 1187, 1979.

Neyman, J., and E. S. Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231 (1933): 289–337.

Niedermeier, Keith E., Norbert L. Kerr, and Lawrence A. Messé. "Jurors' Use of Naked Statistical Evidence: Exploring Bases and Implications of the Wells Effect." *Journal of Personality and Social Psychology* 76, no. 4: 533–42, 1999.

Papineau, David. "The Disvalue of Knowledge." *Synthese* 198, no. 6: 2021.

Pardo, Michael S. “More on the Gettier Problem and Legal Proof:: Unsafe Knowledge Does Not Mean That Knowledge Must Be Safe.” *Legal Theory* 17, no. 1: 75–80, 2011.

———. “Safety vs. Sensitivity: Possible Worlds and the Law of Evidence.” *Legal Theory* 24, no. 1: 50–75, 2018.

———. “The Gettier Problem and Legal Proof.” *Legal Theory* 16, no. 1: 37–57, 2010.

Park, Andrew “Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing.” *UCLA Law Review*, 2019.

Pasquale, Frank. *The Black Box Society: The Secret Algorithms That Control Money and Information.* Harvard University Press, 2016.

Pearson, Karl, and Margaret Moul. “The Problem of Alien Immigration into Great Britian, Illustrated by an Examination of Russian and Polish Jewish Children," *Annals of Eugenics* 1, no. 1: 5–54, 1925.

People v. Johnson, No. 317 Ill. 430 (Illinois Supreme Court 1925).

Perez Carballo, Alejandro. “Good Questions.” In Epistemic Consequentialism, edited by Kristoffer Ahlstrom-Vij and Jeffrey Stewart Dunn. Oxford New York (N.Y.): Oxford university press, 2018.

Pritchard, Duncan. “Legal Risk, Legal Evidence and the Arithmetic of Criminal Justice.” *Jurisprudence* 9, no. 1: 108–19, 2018.

Rai, Arun. “Explainable AI: From Black Box to Glass Box.” *Journal of the Academy of Marketing Science 48*, no. 1: 137–41, 1010.

Rawls, John. *A Theory of Justice.* Harvard University Press, 2009.

re *Winship*, No. 397 (U.S. 358 1970).

Ross, Jacob, and Mark Schroeder. “Belief, Credence, and Pragmatic Encroachment, *Philosophy and Phenomenological Research* 88, no. 2: 259–88, 2014.

Rudner, Richard. "The Scientist Qua Scientist Makes Value Judgments." Philosophy of Science 20, no. 1 (1953): 1–6.

"Rule 404. Character Evidence; Other Crimes, Wrongs, or Acts | Federal Rules of Evidence | LII / Legal Information Institute."

Rysiew, Patrick. "The Context-Sensitivity of Knowledge Attributions." *Nous* 35, no. 4: 477–514, 2001.

Schroeder, Mark. "Stakes, Withholding, and Pragmatic Encroachment on Knowledge." *Philosophical Studies* 160, no. 2: 265–85, 2012.

Schwikkard, Pamela-Jane. *Presumption of Innocence*, Juta, 1999.

Selbst, Andrew D., and Solon Barocas. "The Intuitive Appeal of Explainable Machines." *SSRN Electronic Journal*, 2018.

Shevlin, Henry, and Marta Halina. "Apply Rich Psychological Terms in AI with Care." *Nature Machine Intelligence* 1, no. 4: 165–67, 2019.

Sidgwick, Henry. *The Methods of Ethics.* Cambridge: Cambridge University Press, 2011.

Skeem, Jennifer L., and Christopher T. Lowenkamp. "Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*." *Criminology* 54, no. 4: 680–712, 2016.

Sorensen, Roy. "Future Law: Prepunishment and the Causal Theory of Verdicts." *Noûs* 40, no. 1: 166–83, 2016.

Spanos, Aris. "Is Frequentist Testing Vulnerable to the Base-Rate Fallacy?" Philosophy of Science 77, no. 4 (October 2010): 565–83. https://doi.org/10.1086/656009.

Stanley, Jason. *Knowledge and Practical Interests.* Oxford University Press, 2005.

Stegenga, Jacob. *Medical Nihilism*, Oxford University Press, 2018.

Steinberg, Matthew P. "ESSA Provides Fresh Impetus to Amend Recently Minted Teacher Evaluation Systems or Revise Them Altogether, Armed with This Practical Guidance for Policymakers and School Leaders," National Association of State Boards of Education, 2016.

Strawson, Peter, Fischer, John Martin, and Mark Ravizza, eds. "Freedom and Resentment." In Perspectives on Moral Responsibility, 45–66. Cornell University Press, 2019.

Tadros, Victor. "Rethinking the Presumption of Innocence." *Criminal Law and Philosophy* 1, no. 2: 193–213, 2007.

Tanford, J. Alexander. "The Law and Psychology of Jury Instructions." *Nebraska Law Review* 69, no. 1, 1990

Tomasi, Dardo, and Nora D. Volkow. "Laterality Patterns of Brain Functional Connectivity: Gender Effects." *Cerebral Cortex*, 22, no. 6: 1455–62, 2012.

Unger, Peter. "Philosophical Relativity." *The Philosophical Quarterly* 35, no. 139, : 207, 1985.

Veatch, Henry. "Carl G. Hempel Aspects of Scientific Explanation and Other Essays in the Philosophy of Science . New York: The Free Press, 1965. 505 Pp." Philosophy of Science 37, no. 2 (June 1970): 312–14. https://doi.org/10.1086/288305.

Vitopoulos, Nina A., Michele Peterson-Badali, Shelley Brown, and Tracey A. Skilling. "The Relationship Between Trauma, Recidivism Risk, and Reoffending in Male and Female Juvenile Offenders." *Journal of Child & Adolescent Trauma* 12, no. 3: 351–64, 2019.

Vredenburgh, Kate. "The Right to Explanation*." *Journal of Political Philosophy* 30, no. 2: 09–29, 2022

Wakefield, Jon. "Bayes Factors for Genome-Wide Association Studies: Comparison with P - Values." Genetic Epidemiology 33, no. 1 (January 2009): 79–86. https://doi.org/10.1002/gepi.20359.

Wang, Jackie. *Carceral Capitalism.* Cambridge: Semiotexte/Smart Art, 2018.

Warner, Richard, and Robert H. Sloan. "Making Artificial Intelligence Transparent: Fairness and the Problem of Proxy Variables." *Criminal Justice Ethics* 40, no. 1: 23–39, 2021.

Wasserstein, Ronald L., and Nicole A. Lazar. "The ASA Statement on p -Values: Context, Process, and Purpose." *The American Statistician* 70, no. 2: 129–33, 2016.

Weatherson, Brian. "Knowedge, Bets, and Interests." In *Knowledge Ascriptions*, edited by Jessica Brown and Mikkel Gerken. Oxford: Oxford University Press, 2012.

Weatherson, Brian. "Defending Interest-Relative Invariantism" *Logos & Episteme* 2, no. 4, 591–609, 2011.

Whitman, James Q. *The Origins of Reasonable Doubt: Theological Roots of the Criminal Trial.* New Haven: Yale University Press, 2016.

Williamson, Timothy. *Knowledge and Its Limits.* Oxford: Oxford University Press, 2009.

Winkler, Julia K., Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, et al. "Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition." *JAMA Dermatology* 155, no. 10: 1135–41, 2019