Challenges and examples of rigorous deductive reasoning about socially-relevant issues

Dustin Wehr*

June 12, 2014

Abstract

The most important problems for society are vague, subjective, and filled with uncertainty. The aim of this paper is to help revive the view that giving non-trivial, rigorous deductive arguments concerning such problems –*without* eliminating the complications of vagueness, subjectivity, and uncertainty– is, though very difficult, not problematic in principle, does not require the invention of new logics (classical first order logic will do, at least for a while!), and is something that more of us should be pursuing.

Examples are the main offering of this paper; the rest is an assembly of old and elementary ideas.[1] All are original arguments. One is fully formalized, verified by a first-order theorem prover, and viewable in a readable dynamic HTML format online,[2] and two are mostly-formalized and presented inside. Each is or can be written as a *vaguely-interpreted formal proof*, which is a simple presentation convention introduced in this paper –and used without name since the discovery of predicate logic– for attaching one's vague and subjective intended semantics to the symbols of one's proof, and which is well-suited for the kinds of what-do-you-mean-by? semantic criticisms that are fundamental to arguing about vague and subjective issues. Sketches of some such criticisms for the examples are included.

The first example inside concerns the Berkeley gender bias lawsuit from 1973; it is used as an example in a number of introductory statistics textbooks. The second example inside concerns a murder conviction in Canada that is currently being appealed by the Association in Defense of the Wrongly Convicted.[3] The third example, available online, concerns a case about assisted suicide ruled on by the Supreme Court of Canada in 1993.

*University of Toronto, Department of Computer Science

[1] For us; not so for the people who have expressed the strongest interest in this project, who mostly work in non-mathematical fields. An important goal for me, which I probably won't have touched until after the publication of my dissertation, is to design an unintimidating web-based system where intelligent people without a background in logic can contribute to formal-logic-backed arguments.

[2] http://www.cs.toronto.edu/~wehr/research_docs/sue_rodriguez.html

[3] Thanks to Joanne McLean, Deryck Ramcharitar, and James Lockyer for making the case files available to me.

Contents

1 Introduction

Gottfried Leibniz had a dream, long before the formalization of predicate logic, that some day the rigour of mathematics would find much broader use.

> *"It is true that in the past I planned a new way of calculating suitable for matters which have nothing in common with mathematics, and if this kind of logic were put into practice, every reasoning, even probabilistic ones, would be like that of the mathematician: if need be, the lesser minds which had application and good will could, if not accompany the greatest minds, then at least follow them. For one could always say: let us calculate, and judge correctly through this, as much as the data and reason can provide us with the means for it. But I do not know if I will ever be in a position to carry out such a project, which requires more than one hand; and it even seems that mankind is still not mature enough to lay claim*

to the advantages which this method could provide."

(Gottfried Leibniz, 1706[4])

Unfortunately he was right about the maturity of mankind, and probably still is. This paper is not for cynics; do not bother with it if you're unwilling to at least pretend that mathematicians, logicians, and philosophers can do something to make mankind more reasonable.

What is clear to many of us –that formal logic is *in principle* applicable to arguments about social, contentious, emotionally charged issues– sounds absurd to most people, even highly educated people. The first, rather unambitious goal of this project, is to illustrate this understanding. The second goal, a very difficult and lonely one, is to investigate whether such use of rigorous deduction is worth doing, even if only in our spare time.

There are thousands and thousands of pages by hundreds of scholars that are tangentially related to this work; papers about vagueness in the abstract,[5] the theoretical foundations of Bayesian reasoning,[6] *abstract dialog systems* [7], etc. There is a huge amount of scholarly work on systems and tools and consideration of the theoretically-interesting corner cases, but too little serious work in which the problems take precedence over the tools used to work on them. This paper introduces a project that is the latter kind; we work on important specific problems, attacking general theoretical problems only as-necessary. I am still hopeful that work on argumentation systems will turn out useful, but see the short Section 5 for my reservations about trying to integrate it too early.

Surprisingly, it is the (normative side of the) field of Informal Logic that is probably most related to this project [12][11]. For a long time now they have understood that dialogue-like interactions, or something similar, are essential for arguing about the problems we are concerned with here (Section 2). But I think formal logic has something to contribute here; there are too many examples where good, intelligent scientists and statisticians are given a voice on such problems, only to fail to adhere to the same standards of rigour that they follow in their professional work.[7]

There are commonalities between typical mathematics proofs and proofs about subjective and vague concepts. For example, in both domains, we only need to axiomatize the structures we are thinking about precisely enough for the proof to go through; our proofs about numbers and sets never[8] require complete characterizations, and similarly, for proofs about people, laws, moral values, etc, there is no need to fully *eliminate* vagueness. This is related to the idea of a top-down proof [9], which is, as far as I can tell, the only option when

[4]From translation of a letter to Sophia of Hanover [4]

[5]See [9], where the approach I take to reasoning in the presence of vagueness does not appear to be covered. It could be called *vagueness as plurality of intended models.*

[6]I recommend [6].

[7][8] provides a good example. There Sesardic, a philosopher, contradicts the hasty conclusions of some very reputable statisticians, essentially by applying the same Bayesian quantitative argument, but with much more care taken in constraining the values of the prior probabilities.

[8]Except for proofs about finite structures, I suppose.

[9]Just for fun/to show off the dynamic HTML output, here are two examples of top-down proofs that I wrote while debugging the current system (The second is fully formally verified, the first nearly so.):
5 color theorem: http://www.cs.toronto.edu/~wehr/research_docs/5colortheorem.html

reasoning faithfully about vague concepts. For the remainder of this introduction, I will talk about the differences in reasoning between the two domains. There are three aspects of contentious socially-relevant questions that distinguish them from typical mathematics problems: vagueness[10], subjectiveness[11], and uncertainty. None of these can be eliminated completely without changing the fundamental nature of the problems.

With mathematics problems we can *usually* axiomatize structures sufficiently-precisely at the beginning of our reasoning (when a new area of mathematics is developing), whereas in reasoning about social issues one must delay the precisifying of vague definitions until necessary – in particular, until critics of one's argument are too unclear about one's informal semantics for a symbol to be able to evaluate whether they should accept or reject an axiom that depends on that symbol (this is called a *semantic criticism* in Section 3). To understand the approach to formalizing vagueness used in this project, in which vagueness is represented by plurality of models, it is vital to think of models and the satisfaction relation as the fundamental logical concepts, as opposed to validity. Section 3 should make that clear. See the footnote at the end of Appendix A for consideration of a Sorites problem. Finally, of course questions about vague concepts cannot always be answered in a particular way. What may happen is that the question has different answers depending on how it is precisified, which is determined by the author of an argument that purports to answer that question (and sometimes, indirectly, by the critics of the argument). An illustrative example of this can be made with Newcomb's Paradox[12]; for every English presentation of the problem that I've seen, it is not hard to give two reasonable formalizations that yield opposite answers.

As with vagueness, subjectiveness demands some system of interaction between people on the two sides of an argument, and I am working on an implementation of such a system now. Initially I thought that a dialogue system, with rules designed to ensure progress under certain assumptions, would be essential to move forward with this project. However, my concern that asking mathematically-inclined people to commit to a dialogue, in addition to asking them to reason faithfully about complicated inelegant structures, is asking too much, has led me to (at least temporarily) shift to a simple and lax model of interaction where (a) each proof is owned by an author, and can be critiqued by others; and (b) the author, or a new author, may respond to a critique by modifying their proof or making a new one.[13] A more sophisticated dialogue system will appear in full in my dissertation. Initially I had planned not to include anything about it in this paper, believing it unsatisfactory and too complicated to be practical, but after the helpful criticisms of the anonymous reviewers, I decided to add a brief and informal sketch in Appendix A, in the hope of getting useful feedback. As with vagueness, of course I do not mean to suggest that formal logic can help two parties with conflicting beliefs come to the same answer on, say, questions of ethics. However, where formal logic can help is to find *fundamental* sources of disagreement starting from disagreement on some complex question (which is progress!).

Infinitely-many primes: http://www.cs.toronto.edu/~wehr/research_docs/infinitely-many_primes.html

[10] Classic examples are vague predicates expressing tallness, or baldness

[11] E.g. the weights of various principles of morality

[12] Start at the Wikipedia page if you haven't heard of this and are curious.

[13] For now. This might change, in which case work on argumentation systems becomes especially relevant.

Uncertainty, as I expect many readers of this paper will already agree, is the most difficult of the three complications. Sparsity of information can make it impossible to give an *absolutely*-strong deductive argument for or against a given proposition. But interaction is useful here, too. For example, in Section 7 I give a proof that a key piece of evidence that was used to convict a man of murder has no inculpatory value. Now, I cannot say that the assumptions from which that conclusion (the proposition named ⟨the newspaper hair evidence is neutral or exculpatory⟩) follows are *absolutely* easy to accept, but I confidently challenge anyone to come up with a proof of a strong negation of that conclusion[14] from equally-easy-to-accept assumptions. Hence, I am claiming that my assumptions *are* easy to accept *relative* to what my opponents could come up with.

I use a superficial extension of classical FOL in this project, and *for this particular project* –rigorous deductive reasoning about the kind of questions described in Section 2– that appears to be the right fit. It is vital that the interface between syntax and semantics is as simple as possible, and classical FOL with Tarski semantics is the best in this respect. That said, in this paper, which is long enough already, I do not give the argument the space it deserves. In my dissertation, I take some space to explain how common forms of defeasible reasoning can be carried out in deductive logic, and how ideas from e.g. modal or epistemic or temporal logic can be used as-necessary without needing to build them into the definition of the logic (complicating the semantics). On the other hand, it may turn out that the more-concise syntax of some such such logic is often enough helpful for reading and writing proofs, without obscuring the meaning, that it would be wise to adopt it for the system. That, I think, should be decided after a great deal more experience.

In Section 2, I clarify what I mean by "social, contentious, emotionally charged issues." In Section 3 I introduce the concept of a *vaguely-interpreted formal proof*, which amounts to a controlled way of pairing a formal proof with its intended semantics. Though very simple, I find the concept extremely useful for precluding uses of formal logic that exploit vagueness and impose too great a chore on potential critics (who have no option but to guess the author's intended semantics, likely needing to consider a wide range of possible intended semantics).[15] Section 6 contains relatively simple vaguely-interpreted proofs about a famous lawsuit against U.C. Berkeley for gender discrimination in their admissions process. In Section 7 I give an unusually rigorous Bayesian argument about the weight of a piece of key evidence in a murder case that is currently being appealed by the Association in Defense of the Wrongly Convicted.

[14] e.g. that the likelihood ratio is much larger than 1.

[15] An example of this is Gödel's posthumously published ontological proof[3]. In my dissertation, I will demonstrate an attempt to turn it into a vaguely-interpreted proof that comes close to having the *locally refutable* property described at the end of Section 3, with the goal of convincing the reader that doing so is not possible.

2 Problem domain

Provided the uncertainty involved in a problem is not too great, or that it *is* too great but one side of the argument has the burden of proof, it is my view, from over two years of working on this project, that the main impediments to rigorous deductive reasoning about socially relevant issues are *a*) conventional mathematical modeling difficulty; *b*) conventional mathematical problem solving difficulty[16]; and *c*) tedium[17]. These are strong impediments. For that reason, I think it is worthwhile to describe the questions that I think are best-suited for rigorous deductive reasoning. These are *contentious questions with ample time available*. Typical sources of such problems are public policy and law. Without ample time, it may be detrimental to insist on deductive reasoning; as pointed out in many places, when complete heuristic reasoning and incomplete deductive reasoning are the only options, it is probably best to go with the former. Without contentiousness, there is little motive for employing fallacious reasoning and rhetoric to advance one's position, and this, I think, defeats much of the benefit of using formal logic (or some approximation of it, as appears in mathematics journals). At the same time, lack of contentiousness does not proportionally reduce the work required for rigour, so we are left with less expected benefit relative to cost. Leibniz was conscious of this point:

> *I certainly believe that it is useful to depart from rigorous demonstration in geometry because errors are easily avoided there, but in metaphysical and ethical matters I think we should follow the greatest rigor, since error is very easy here. Yet if we had an established characteristic*[18] *we might reason as safely in metaphysics as in mathematics.*
>
> (Gottfried Leibniz, 1678, Letter to Walter von Tschirnhaus[5])

In contrast, some prominent Logical Positivists seem to have thought that this is not a crucial constraint (e.g. Hans Reichenbach's work on axiomatizing the theory of relativity).

[16]Two of my proofs (one of which, the Leighton Hay argument, appears in this paper) are currently contingent on the truth of mathematical statements that I cannot easily prove. This is my attitude about such statements: there are mathematicians out there who can easily prove or disprove them, but I think it would be premature to call upon them until proofs of the statements have actually been demanded by critics (called a *mathematics detail criticism* in the paper). In the meantime, I give some empirical evidence of their truth (in this case, numerical evaluation of a complicated integral, without error bounds), and there are other, more-subjective axioms of the proof that are much easier targets for criticism. It may even be wise to build in some precedence in the rules for criticizing a vaguely interpreted proof, whereby under certain conditions (which aren't obvious to me at present) one must accept the axioms that involve vague and subjective concepts before demanding a proof of a purely-mathematical claim (of course, one should always be able to present disproofs).

[17]This has been the hardest of the three for me to cope with. My hope is that this impediment will be reduced by making the construction of such arguments a collaborative, social process on the web.

[18]Leibniz is referring to the practical system/method that he envisioned, but was unable to devise.

3 Vaguely-interpreted formal proofs

In this section I am just giving a name to a kind of document that most teachers of first-order logic have used at least implicitly. The point is just to make concrete and explicit a bridge between the formal and informal, providing a particular way, that is amenable to criticism, for an author of a proof to describe their intended semantics in the metalanguage.

The definition of *vaguely-interpreted formal proof* is tailored for classical many-sorted first order logic, but it will be clear that a similar definition can be given for any logic that has a somewhat Tarski-like semantics, including the usual untyped classical first order logic, or fancier versions of many-sorted first-order logic.[19] A very minor extension of the usual definition of many-sorted first order logic (where sorts must be interpreted as disjoint sets that partition the universe) is used here and in the examples: it includes easily-eliminable sort operators and partial functions/undefinedness, the latter based on [2]. A language is just a set of symbols, each of which is designated a constant, predicate, function, sort, or sort-operator symbol. A signature is a language together with an assignment of types to the symbols (or, in the case of sort operators, just an assignment of arities).

There are four kinds of formal axioms that appear in a vaguely-interpreted formal proof:

- An assumption imposes a significant constraint on the semantics of vague symbols (most symbols other than common mathematical ones), even when the semantics of the mathematical symbols are completely fixed.
- A claim does not impose a significant constraint on the semantics of vague symbols. It is a proposition that the author of the proof is claiming would be formally provable upon adding sufficiently-many uncontroversial axioms to the theory.
- A simplifying assumption is a special kind of an assumption, although what counts as a simplifying assumption is vague. The author of the proof uses it in the same way as in the sciences; it is an assumption that implies an acknowledged inaccuracy, or other technically-unjustifiable constraint, that is useful for the sake of saving time in the argument, and that the author believes does not bias the results.
- A definition is, as usual, an axiom that completely determines the interpretation of a new symbol in terms of the interpretations of previously-introduced symbols.

A language interpretation guide g for (the language of) a given signature is simply a function that maps each symbol in the language to a chunk of natural language text, which describes, often vaguely, what the author intends to count as *an* intended interpretation of the symbol; due to the vagueness in the problems we are interested in, a set of axioms will have many intended models. Typically $g(s)$ will be between the length of a sentence and a long paragraph, but can be longer.

A signature's language has *sort symbols*, which structures must interpret as disjoint subsets of the universe. A language can also have *sort operator symbols*, which are second order function symbols that can only be applied to sorts. In this project sort operators have a

[19] An earlier version of this paper included the syntax and semantics of such a fancier logic. That logic is a little more convenient for formalization, but it does not contribute to the goals of the paper.

nonvital role, used for uniformly assigning names and meanings to sorts that are definable as a function of simpler sorts, when that function is used multiple times and/or is applied to vague sorts (i.e. sorts in $\mathcal{L}_{\mathsf{vague}}$, introduced below).[20] A signature assigns sorts to its constants, and types to its function and predicate symbols. In this project, types are mostly used as a lightweight way of formally restricting the domain on which the informal semantics of a symbol must be given (by the language interpretation guide). To see why they are beneficial, suppose that we didn't have them, e.g. that we were using normal FOL. For the sake of clarity, we would nonetheless usually need to specify types either informally in the language interpretation guide, or formally as axioms. In the first case, we inflate the entries of the language interpretation guide with text that rarely needs to be changed as an argument progresses, and that often can be remembered sufficiently after reading it only once. In the second case, we clutter the set of interesting axioms (e.g. the non-obvious and controversial axioms) with uninteresting typing axioms.

A sentence label is one of $\{\mathsf{assum}, \mathsf{simp}, \mathsf{claim}, \mathsf{defn}, \mathsf{goal}\}$, where assum is short for *assumption* and simp is short for *simplifying assumption*. A symbol label is one of $\{\mathsf{vague}, \mathsf{math}, \mathsf{def}\}$.

A vaguely-interpreted formal proof is given by

- A signature Σ.
- A set of well-typed Σ-sentences Γ called the *axioms*.
- An assignment of symbol labels to the symbols of Σ. If $\mathcal{L}$ is the language of Σ, then for each symbol label l we write $\mathcal{L}_l$ for the symbols assigned label l.
- An assignment of sentence labels to the elements of Γ, with one sentenced label goal. For each sentence label l we write Γ_l for the sentences in Γ labeled l.
- An assignment of one of the sentence labels assum or simp to each type assignment of Σ. These typing declarations can be viewed as sentences too, and though they will usually be regular assumptions (labeled assum), occasionally it's useful to make one a simplifying assumption (labeled simp).
- The sentences in Γ_{defn} define the constant, function, and predicate symbols in $\mathcal{L}_{\mathsf{def}}$. Function and predicate symbol definitions have a form like $\forall x_1{:}S_1. \ldots. \forall x_k{:}S_k.\ f(x_1, \ldots, x_k) = t$ where t can be a term or formula (in the latter case, replace $=$ with $\leftrightarrow$) and the S_i are sorts.
- $\mathcal{L}_{\mathsf{vague}}, \mathcal{L}_{\mathsf{math}}, \mathcal{L}_{\mathsf{def}}$ are disjoint languages, $\mathcal{L}_{\mathsf{vague}}$ does not contain any sort-operator symbols,[21] and $\mathcal{L}_{\mathsf{def}}$ contains neither sort nor sort-operator symbols[22].
- g is a language interpretation guide for a subset of the language of Σ that includes $\mathcal{L}_{\mathsf{vague}}$ and $\mathcal{L}_{\mathsf{math}}$. So, giving explicit informal semantics for a defined symbol is optional.
- Γ_{goal} is provable from $\Gamma_{\mathsf{assum}} \cup \Gamma_{\mathsf{simp}} \cup \Gamma_{\mathsf{claim}} \cup \Gamma_{\mathsf{defn}}$.

[20] For example, if our proof only needs the power set of one mathematical sort S (in $\mathcal{L}_{\mathsf{math}}$), then using a sort operator would have little benefit over just introducing another mathematical sort symbol named 2^S. I cannot say the same if S is a vague sort (in $\mathcal{L}_{\mathsf{vague}}$), since then we would have to introduce 2^S as a vague sort as well, and I think minimizing the number of vague symbols is usually desirable.

[21] I suppose that restriction could be lifted, but I haven't had any desire for vague sort operators in all the time I've worked on this project.

[22] Another inessential constraint, which I've added simply so that I don't have to include something in the grammar for defining sorts or sort-operators in terms of other sorts and sort operators

- For each $\psi \in \Gamma_{\mathsf{claim}}$, any reader in the intended audience of the proof can come up with a set of $\mathcal{L}_{\mathsf{math}}$-sentences Δ, which are true with respect to the (informal) semantics given by g, such that $\Gamma_{\mathsf{assum}} \cup \Gamma_{\mathsf{defn}} \cup \Gamma_{\mathsf{simp}} \cup \Delta$ proves ψ. See following paragraph for a more-precise condition.

$\mathcal{L}_{\mathsf{math}}$ is intended to be used mostly for established mathematical structures, but in general for structures that both sides of an argument agree upon sufficiently well that they are *effectively objective with respect to* Γ_{claim}. For each person p in the intended audience of the proof, let Δ_p be the set of $\mathcal{L}_{\mathsf{math}}$-sentences that p can eventually and permanently recognize as true with respect to the informal semantics given by g. Then we should have that $\bigcap_{p \in \text{audience}} \Delta_p$ is consistent and when combined with $\Gamma_{\mathsf{assum}} \cup \Gamma_{\mathsf{defn}} \cup \Gamma_{\mathsf{simp}}$ proves every claim in Γ_{claim}. If that is not the case, then there is some symbol in $\mathcal{L}_{\mathsf{math}}$ that should be in $\mathcal{L}_{\mathsf{vague}}$, or else the intended audience is too broad.

The purpose of the language interpretation guide is for the author to convey to readers what they consider to be an acceptable interpretation of the language. Subjectiveness results in different readers interpreting the language differently, and vagueness results in each reader having multiple interpretations that are acceptable to them. Nonetheless, an ideal language interpretation guide is detailed enough that readers will be able to conceive of a vague set of *personal* Σ*-structures* that is precise enough for them to be able to accept or reject each assumption (element of $\Gamma_{\mathsf{assum}} \cup \Gamma_{\mathsf{simp}}$) independent of the other axioms. When that is not the case, the reader should raise a semantic criticism (defined below), which is similar to asking "What do you mean by X?" in natural language.

In more detail, to review a vaguely-interpreted proof π with signature Σ and language $\mathcal{L}$, you read the language interpretation guide g, and the axioms Γ, and either accept π or criticize it in one of the following ways:

(1) Semantic criticism: Give $\phi \in \Gamma$ and at least one symbol s of $\mathcal{L}_{\mathsf{vague}}$ that occurs in ϕ, and report that $g(s)$ is not clear enough for you to *evaluate* ϕ, which means to conclude that all, some, or none of your personal Σ-structures satisfy ϕ. If you cannot resolve this criticism using communication with the author in the metalanguage, then you should submit a Σ-sentence ψ to the author, which is interpreted by the author as the question: Is ψ consistent with g?

(2) Rigor criticism: Criticize the inclusion of a symbol in $\mathcal{L}_{\mathsf{math}}$, or do the same as in (1) but for $\mathcal{L}_{\mathsf{math}}$. This is the mechanism by which one can insist that vague terms be recognized as such. The same can be done when ϕ is a type assignment or sort constraint, in which case ψ is a Σ-sentence that uses sort symbols as unary predicate symbols.

(3) Mathematics detail criticism: Ask for some claim in Γ_{claim} to be proved from simpler claims (about $\mathcal{L}_{\mathsf{math}}$ interpretations).

(4) Subjective criticism: Reject some sentence $\phi \in \Gamma_{\mathsf{assum}} \cup \Gamma_{\mathsf{simp}}$, which means to conclude that at least one of your personal $\mathcal{L}$-structures falsifies ϕ. If you wish to communicate this to the author, you should additionally communicate one of the following:

(a) Tentative commitment to $\neg\phi$, i.e. conclude that all of your personal Σ-structures falsify ϕ.

(b) Tentative commitment to the independence of ϕ, i.e. conclude that ϕ is also satisfied by at least one of your personal Σ-structures. Intuitively, this means that ϕ corresponds to a *simplifying assumption* that you are not willing to adopt.

In the context of its intended audience, we say that a vaguely-interpreted formal proof is locally-refutable if no member of the intended audience raises semantic or rigor criticisms when reviewing it. A locally-refutable proof has the desirable property that by using the language interpretation guide g, any member of the audience can evaluate each of the axioms of the proof independently of the other axioms. Local-refutability is the ideal for vaguely-interpreted formal proofs. It is a property that is strongly lacking –and not strived for– in most mathematical arguments in economics or game theory, for example, and in every sophisticated abuse of statistics. When an argument is far from locally-refutable, it is hard to criticize in a standard, methodical manner, and that ease of criticism is a central goal of this project.

4 The role of formal logic

Remove formal logic from this project and there is no benefit over our current system of arguing with each other through papers and blog posts and shouting. Those mediums are easier to work in, and superior if one is interested in persuasion; any vaguely-interpreted formal proof can be made more persuasive if converted to an informal argument that mixes natural language and mathematics in the normal way we use in conference and journal papers. The problem with that, from the point of view of this project, is that unsound, invalid, misleading, unfair, and otherwise bad arguments benefit from the lax regulations as much *or more than* good arguments do. This project uses deductive formal logic because it is our best tool for forcing the weaknesses of arguments to be exposed. Thus **the role of formal logic in this project is regulatory, and nothing more than that.** The success of this project rides on the regulatory benefit outweighing the overhead of formalization.

I have caught omissions in my own reasoning thanks to the constraints of formal deductive logic, things that never occurred to me in thinking and talking about an issue for years. I have gained respect for my opponents on *every* issue I've attempted arguments about, forced to consider the issue in its full complexity (e.g. Canada's lifetime ban against blood donations from men who have had sex with men,[23] assisted suicide in Canada, the evidence for anthropogenic global warming[24]). It is hard to write a person off as stupid, relative to

[23]If the ban is lifted, is there non-negligible probability of a significant increase in the rate of people lying in the self-disclosure part of the system? Reasoning deductively, one *must* consider this non-obvious question, and I have found no way to derive my target conclusion (that, with additional safeguards, the ban should be lifted) without making an assumption that is not far from explicitly answering the question 'no'.

[24]The closest thing we have to a strong deductive argument, that I have found, comes from the Berkeley Earth Surface Temperature project, which has mostly been ridiculed by climate science researchers, who simply view it as making no significant advancement in climate science, ignoring or not valuing the fact that

oneself, after a great struggle to find acceptable formal assumptions from which it follows logically that they are wrong.

5 Example: Sue Rodriguez at the Supreme Court of Canada

This example is fully formalized, and verified valid by reduction to (easy) FOL validity problems that were solved by CVC4 and Vampire via System on TPTP[25]. It can be read here:

http://www.cs.toronto.edu/~wehr/research_docs/sue_rodriguez.html

Following the proof, there is a short section of some sample criticism. Eventually you will be able to edit any part of the proof (making a copy linked to the original, with the author of the original notified of any changes you make, or concurrently editing the original if given appropriate permissions by the author[26]), and there will be a built-in feature for attaching criticisms of the type described toward the end of Section 3 to individual axioms.

The rules for criticizing vaguely-interpreted formal proofs, and the definition of "vaguely-interpreted proofs" itself, will evolve as I and (probably a very small number of) others gain experience using the system and criticizing each others' arguments. That has been the highly-effective approach taken by the curators of Wikipedia, and I believe strongly that it is the right approach to take for this project as well.

6 Example: Berkeley gender bias lawsuit

The following table summarizes UC Berkeley's Fall 1973 admissions data for its six largest departments. Across all six departments, the acceptance rates for men and women are about 44.5% and 30.4% respectively. The large observed bias prompted a lawsuit against the university, alleging gender discrimination.[27] In [1] it was argued that the observed bias was actually due to a tendency of women to disproportionately apply to departments that have high rejection rates for both sexes.

---

it seeks to minimize the use of argument from expert opinion.

[25] www.cs.miami.edu/~tptp/cgi-bin/SystemOnTPTP

[26] All of this functionality is conveniently provided by Google's Realtime API.

[27] The data given is apparently the only data that has been made public. The lawsuit was based on the data from all 101 graduate departments, which showed a pattern similar to what the data from the 6 largest shows.

	Male		Female		Total	
Department	Applied	Accepted	Applied	Accepted	Applied	Accepted
D_1	825	512 (62%)	108	89 (82%)	933	601 (64%)
D_2	560	353 (63%)	25	17 (68%)	585	370 (63%)
D_3	325	120 (37%)	593	202 (34%)	918	322 (35%)
D_4	417	138 (33%)	375	131 (35%)	792	269 (34%)
D_5	191	53 (28%)	393	94 (24%)	584	147 (25%)
D_6	373	22 (6%)	341	24 (7%)	714	46 (6%)

The first argument I give is similar to the final analysis given in [1],[28] though it makes weaker assumptions (Assumption 2 in particular: their corresponding, implicit assumption is obtained by replacing the parameters .037 and 9 with 0s). The argument resolves the apparent paradox by assuming a sufficiently-precise definition of "gender discrimination" and reasoning from there. More precisely, it first fixes a definition of "gender discrimination", and then defines (in natural language) a hypothetical admissions protocol that prevents gender discrimination by design. Considering then a hypothetical round-of-admissions scenario that has the same set of applications as in the actual round of admissions, if we assume that the ungendered departmental acceptance rates are not much different in the hypothetical scenario, then it can be shown that the overall bias is actually *worse* for women in the hypothetical scenario. Since the hypothetical scenario has no gender discrimination by design, and is otherwise very similar to the real scenario, we conclude that the observed bias cannot be blamed on gender discrimination.

The second argument tells us why it is that our vagueness about "gender discrimination" resulted in an apparent paradox; namely, we were implicitly admitting definitions of "gender discrimination" that allow for the question of the presence/absence of discrimination to depend on whether or not the sexes apply to different departments at different rates. If we forbid such definitions, then to prove that the gendered departmental acceptance rates *do not* constitute gender discrimination, it should suffice to show that there is an overall bias *in favour* of women in any hypothetical admissions round in which the gendered departmental acceptance rates are close to what they actually were, and where men and women apply to each department at close to the same rate.

I'll use g to refer to the *language interpretation guide* for the language $\mathcal{L}$ of this argument. $\mathcal{L}\backslash\mathcal{L}_{\mathsf{math}}$ consists of:

- The constant $\mathsf{Acc}_{\mathsf{hyp}}$.
- The propositional variables (i.e. 0-ary predicate symbols) $\langle$bias only evidence$\rangle$, $\langle$lawsuit should be dismissed$\rangle$, $\langle$gender uncor with ability in each dept$\rangle$.

$\mathcal{L}_{\mathsf{math}}$ consists of:

[28]The paper is written to convey the subtlety of the statistical phenomenon involved (an instance of "Simpson's Paradox"), and so considers several poor choices of statistical analyses before arriving at the final one.

- The constants App, Acc, App^m, App^f, $\mathsf{App}^1, \ldots, \mathsf{App}^6$. Since the elements of these sets are not in the universe, their semantics are determined by axioms that assert their sizes and the sizes of sets formed by intersecting and unioning them with each other. The reader can check that this is fits with the definition given in the paragraph above that introduces vaguely-interpreted formal proofs.
- A number of mathematical symbols that have their standard meaning: constants $0, 1, 512, 825, \ldots$, function symbols $|\cdot|, \cap, \cup, +, -, *, /$, predicate symbols $<, =$.
- The sorts $\mathcal{A}, \mathbb{N}, \mathbb{Q}$ (see below for g's entries for them), with $\mathbb{Q}$ and $\mathcal{A}$ the top-level sorts, and $\mathbb{N} \subseteq \mathbb{Q}$. Recall that the top-level sort symbols must be interpreted as a partition of the universe.

The types of the function/predicate symbols[29] are as follows. With respect to the definition of *vaguely-interpreted formal proof* from Section 3, they are all *assumptions* as opposed to *simplifying assumptions*.

$$\begin{array}{rcl} \mathsf{App}, \mathsf{Acc}, \mathsf{Acc}_{\mathsf{hyp}}, \mathsf{App}^m, \mathsf{App}^f, \mathsf{App}^1, \ldots, \mathsf{App}^6 & : & \mathcal{A} \\ |\cdot| & : & \mathcal{A} \to \mathbb{N} \\ 0, 1, 512, 825, \ldots & : & \mathbb{N} \\ \cap, \cup & : & \mathcal{A} \times \mathcal{A} \to \mathcal{A} \\ +, -, * & : & \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q} \\ / & : & \mathbb{Q} \times \mathbb{Q} \to_? \mathbb{Q}^{29} \\ < & : & \mathbb{Q} \times \mathbb{Q} \to \mathbb{B}^{30} \end{array}$$

6.1 First Argument

The *goal sentence* is the following implication involving propositional variables whose informal meanings, given by the language interpretation guide g, will be given next.

$$\langle \mathsf{gender\ uncor\ with\ ability\ in\ each\ dept} \rangle \wedge \langle \mathsf{bias\ only\ evidence} \rangle \Rightarrow \langle \mathsf{lawsuit\ should\ be\ dismissed} \rangle$$

$g(\langle \mathsf{bias\ only\ evidence} \rangle)$ consists of the above table, and then the assertion: "The bias shown in the data is the only evidence put forward by the group who accused Berkeley of gender discrimination."

$g(\langle \mathsf{gender\ uncor\ with\ ability\ in\ each\ dept} \rangle)$ we take to be just "Assumption 1" from [1], which I quote here:

[29] Besides $=$, which is untyped.

[29] $\to_?$ denotes the type of a partial function. The version of many-sorted FOL I use has build-in (AKA "first-class") partial functions, based on [2].

[30] $\mathbb{B}$ is the type for booleans; technically it is not a sort, so its elements are not in the universe of discourse.

> *Assumption 1 is that in any given discipline male and female applicants do not differ in respect of their intelligence, skill, qualifications, promise, or other attribute deemed legitimately pertinent to their acceptance as students. It is precisely this assumption that makes the study of "sex bias" meaningful, for if we did not hold it any differences in acceptance of applicants by sex could be attributed to differences in their qualifications, promise as scholars, and so on. Theoretically one could test the assumption, for example, by examining presumably unbiased estimators of academic qualification such as Graduate Record Examination scores, undergraduate grade point averages, and so on. There are, however, enormous practical difficulties in this. We therefore predicate our discussion on the validity of assumption 1.* [1]

$g(\langle \mathsf{lawsuit\ should\ be\ dismissed} \rangle)$ = "The judge hearing the suit against Berkeley should dismiss the suit on grounds of lack of evidence."

$g(\mathbb{Q})$ = "The rational numbers."

$g(\mathcal{A})$ = "The powerset of App. Note that the individual applications are not in the universe of discourse (though each singleton set is), since they are not required for the proof."

g also says that

- $0, 1, 512$, etc are the expected numerals.
- $|\cdot|$ is the function that gives the size of each set in $\mathcal{A}$.
- $\cap, \cup, +, -, *$ are the expected binary functions on $\mathcal{A}$ and $\mathbb{Q}$ respectively..
- $/$ is division on $\mathbb{Q}$, which is defined iff the second argument is not 0.
- $<$ is the usual ordering on $\mathbb{Q}$.

Recall that the next 11 symbols are all 0-ary predicate symbols.

$g(\mathsf{App})$ = "App is the set of applications. Its size is 4526 (sum of the entries in the two "Applied" columns of the table)."

$g(\mathsf{Acc})$ = "Acc is the set of (actual) accepted applications. Its size is 1755 (sum of the entries in the two "Accepted" columns of the table)."

$g(\mathsf{Acc}_{\mathsf{hyp}})$ is a fairly long text: "We need a sufficiently-precise, context-specific definition of "gender discrimination", and to get it we imagine a hypothetical scenario. An alternative admissions process is used, which starts with exactly the same set of applications App, and then involves an elaborate[31], manual process of masking the gender on each of them (including any publications and other supporting materials). The application reviewers, while reading the applications and making their decisions, are locked in a room together without access to outside information, except that interviews are done over computer using an instant messaging client (which, of course, is monitored to make sure the gender of

[31]It need not be efficient/economical, since we are only introducing the scenario as a reasoning device.

the applicant remains ambiguous). Then, $\mathsf{Acc}_{\mathsf{hyp}}$ is the set of accepted applications in the hypothetical scenario."

$g(\mathsf{App}^m)$ = "App^m is a subset of App of size 2691 (sum of the first "Applied" column in the table), specifically the applications where the applicant is male."

$g(\mathsf{App}^f)$ = "App^f is a subset of App of size 1835 (sum of the second "Applied" column in the table), specifically the applications where the applicant is female."

For $d = 1, \ldots, 6$:
$g(\mathsf{App}^d)$ = "App^d is the set of applications for admission into department d."

Definition 1. For $g \in \{m, f\}$ and $d \in \{1, \ldots, 6\}$:

$$\mathsf{App} \coloneqq \mathsf{App}^m \uplus \mathsf{App}^f$$

$$\mathsf{App}^{d,g} \coloneqq \mathsf{App}^d \cap \mathsf{App}^g$$

$$\mathsf{Acc}^{d,g} \coloneqq \mathsf{App}^{d,g} \cap \mathsf{Acc}$$

$$\mathsf{Acc}^{d,g}_{\mathsf{hyp}} \coloneqq \mathsf{App}^{d,g} \cap \mathsf{Acc}_{\mathsf{hyp}}$$

Definition 2. For $x, y, z \in \mathbb{Q}$, we write $z \in [x \pm y]$ for $x - y \leqslant z \leqslant x + y$.

Assumption 1. In the hypothetical scenario, the number of applicants of gender g accepted to department d is as close as possible to what we'd expect assuming that gender is uncorrelated with ability within the set of applicants to department d. For $d \in \{1, \ldots, 6\}$ and $g \in \{m, f\}$:

$\langle$gender uncor with ability in each dept$\rangle \Rightarrow$

$$|\mathsf{Acc}^{d,g}_{\mathsf{hyp}}| \in \left[|\mathsf{Acc}^{d}_{\mathsf{hyp}}| \cdot \frac{|\mathsf{App}^{d,g}|}{|\mathsf{App}^{d}|} \pm 1/2 \right]$$

$$\frac{|\mathsf{Acc}^{d,g}_{\mathsf{hyp}}|}{|\mathsf{Acc}^{d}_{\mathsf{hyp}}|} = \frac{|\mathsf{App}^{d,g}|}{|\mathsf{App}^{d}|}$$

Assumption 2. Assuming that gender is uncorrelated with ability within the set of applicants to department d, the number of applicants accepted to department d in the hypothetical scenario is close to the number accepted in the real scenario. That is, the overall, non-gendered departmental acceptance rates do not change much when we switch to gender-blind reviews. We require that a model satisfies at least one of the following two quantifications of that idea. For $d \in \{1, \ldots, 6\}$:

$\langle$gender uncor with ability in each dept$\rangle \Rightarrow$

$$\left(\textstyle\bigwedge_{1 \leqslant d \leqslant 6} |\mathsf{Acc}^{d}| \cdot (1 - .037) \leqslant |\mathsf{Acc}^{d}_{\mathsf{hyp}}| \leqslant |\mathsf{Acc}^{d}| \cdot (1 + .037) \right)$$
$$\vee \quad \left(\textstyle\bigwedge_{1 \leqslant d \leqslant 6} |\mathsf{Acc}^{d}_{\mathsf{hyp}}| \in \left[|\mathsf{Acc}^{d}| \pm 9 \right] \right)$$

To illustrate the first form, the bounds for the departments with the fewest and greatest number of accepted applicants are:

$$45 \leqslant |\mathsf{Acc}^6_{\mathsf{hyp}}| \leqslant 47 \quad \text{and} \quad 579 \leqslant |\mathsf{Acc}^1_{\mathsf{hyp}}| \leqslant 623$$

Definition 3. For $g \in \{m, f\}$:

$$\mathsf{accRate}^g \coloneqq \mathsf{Acc}^g/\mathsf{App}^g \quad \text{and} \quad \mathsf{accRate}^g_{\mathrm{hyp}} \coloneqq \mathsf{Acc}^g_{\mathsf{hyp}}/\mathsf{App}^g$$

Assumption 3. If $\langle$bias only evidence$\rangle$ and

$$\frac{\mathsf{accRate}^m_{\mathrm{hyp}}}{\mathsf{accRate}^f_{\mathrm{hyp}}} > \frac{\mathsf{accRate}^m}{\mathsf{accRate}^f}$$

then $\langle$lawsuit should be dismissed$\rangle$

Simplifying Assumption 1. $\langle$bias only evidence$\rangle$

Claim 1.

$$\langle\text{gender uncor with ability in each dept}\rangle \quad \Rightarrow \quad \frac{\mathsf{accRate}^m_{\mathrm{hyp}}}{\mathsf{accRate}^f_{\mathrm{hyp}}} > \frac{\mathsf{accRate}^m}{\mathsf{accRate}^f}$$

Proof. It is not hard to formulate this as a linear integer programming problem, where the variables are the sizes of the sets $\mathsf{Acc}^{d,g}_{\mathsf{hyp}}$. Coming up with inequalities that express the previous axioms and the data axioms from Section 6.3 is easy. Reduce the Claim itself to a linear inequality, and then negate it. One can then proof using any decent integer programming solver that the resulting system of equations is unsatisfiable. □

Claim 2. The goal sentence easily follows from the previous three propositions.

$\langle$gender uncor with ability in each dept$\rangle \wedge \langle$bias only evidence$\rangle \Rightarrow \langle$lawsuit should be dismissed$\rangle$

6.2 Second argument

This second argument better captures the intuition of the usual informal resolution of the apparent paradox; the observed bias is completely explained by the fact that women favored highly-competitive departments (meaning, with higher rejection rates) more so than men. We show that there is an overall bias *in favour* of women in any hypothetical admissions round in which the gendered departmental acceptance rates are close to what they actually were, and where men and women apply to each department at close to the same rate.

In this argument, the set of applications in the hypothetical scenario can be different from those in the real scenario, so we introduce the new symbols $\mathsf{App}^d_{\mathsf{hyp}} : \mathcal{A}$ for $1 \leqslant d \leqslant 6$.

The hypothetical admissions round is similar to the true admissions round (Axioms 4 and 6) except that men and women apply to each department at close to the same rate (Assumption 5) - meaning the fraction of male applications that go to department d is

close to the fraction of female applications that go to department d. We need to update the language interpretation guide entries $g(\mathsf{App}^d_{\mathsf{hyp}})$ and $g(\mathsf{Acc}_{\mathsf{hyp}})$ to reflect these alternate assumptions.

This proof uses Definitions 1 and 2 from the previous proof.

Assumption 4. In the hypothetical round of admissions, the total number of applications to department d is the same as in the actual round of admissions. Likewise for the total number of applications from men and women.[32]
For $d \in \{1, \ldots, 6\}$ and $g \in \{m, f\}$:

$$|\mathsf{App}^d_{\mathsf{hyp}}| = |\mathsf{App}^d|, \quad |\mathsf{App}^g_{\mathsf{hyp}}| = |\mathsf{App}^g|$$

Assumption 5. In the hypothetical scenario, gendered departmental *application* rates are close to gender-independent. For $d \in \{1, \ldots, 6\}$ and $g \in \{m, f\}$:

$$|\mathsf{App}^{d,g}_{\mathsf{hyp}}| \in \left[|\mathsf{App}^g_{\mathsf{hyp}}| \cdot \frac{|\mathsf{App}^d_{\mathsf{hyp}}|}{|\mathsf{App}_{\mathsf{hyp}}|} \pm 6 \right]$$

Assumption 6. In the hypothetical scenario, gendered departmental *acceptance* rates are close to the same as in the real scenario.
For $d \in \{1, \ldots, 6\}$ and $g \in \{m, f\}$:

$$|\mathsf{Acc}^{d,g}_{\mathsf{hyp}}| \in \left[\frac{|\mathsf{Acc}^{d,g}|}{|\mathsf{App}^{d,g}|} \cdot |\mathsf{App}^{d,g}_{\mathsf{hyp}}| \pm 6 \right]$$

Claim 3. $\mathsf{accRate}^f_{\mathrm{hyp}} > \mathsf{accRate}^m_{\mathrm{hyp}}$

Proof. As in the previous proof, it is easy to reduce this to a linear integer programming problem. Coming up with constraints that express the previous axioms and the data axioms from the next section is easy. Then, add the constraint

$$\left(\sum_{1 \leqslant d \leqslant 6} |\mathsf{Acc}^{d,f}_{\mathsf{hyp}}| \right) / |\mathsf{App}^f| \leqslant \left(\sum_{1 \leqslant d \leqslant 6} |\mathsf{Acc}^{d,m}_{\mathsf{hyp}}| \right) / |\mathsf{App}^m|$$

which expresses the negation of the Claim (recall that $|\mathsf{App}^m|$ and $|\mathsf{App}^f|$ are constants). Finally, prove that the resulting system of equations is unsatisfiable. □

Assumption 7. If $\langle$bias only evidence$\rangle$ and $\mathsf{accRate}^f_{\mathrm{hyp}} > \mathsf{accRate}^m_{\mathrm{hyp}}$ then $\langle$lawsuit should be dismissed$\rangle$

Simplifying Assumption 1 from the previous proof, which just asserts $\langle$bias only evidence$\rangle$, is also used here. From it, Assumption 7, and Claim 3, the goal sentence $\langle$lawsuit should be dismissed$\rangle$ follows immediately.

[32]This axiom could be weakened in principle, by replacing the equations with bounds, but doing so in the obvious way introduces nonlinear constraints, and then I would need to use a different constraint solver.

6.3 Data Axioms

Assumption 8.

$$|\mathsf{App}| = 4526, \quad \bigwedge_{1 \leqslant d \leqslant 6} \mathsf{App}^d \subseteq \mathsf{App}, \quad \mathsf{Acc} \subseteq \mathsf{App}, \quad \mathsf{Acc}_{\mathsf{hyp}} \subseteq \mathsf{App}$$

$$|\mathsf{App}^{1,m}| = 825, \quad |\mathsf{Acc}^{1,m}| = 512, \quad |\mathsf{App}^{1,f}| = 108, \quad |\mathsf{Acc}^{1,f}| = 89$$
$$|\mathsf{App}^{2,m}| = 560, \quad |\mathsf{Acc}^{2,m}| = 353, \quad |\mathsf{App}^{2,f}| = 25, \quad |\mathsf{Acc}^{2,f}| = 17$$
$$|\mathsf{App}^{3,m}| = 325, \quad |\mathsf{Acc}^{3,m}| = 120, \quad |\mathsf{App}^{3,f}| = 593, \quad |\mathsf{Acc}^{3,f}| = 202$$
$$|\mathsf{App}^{4,m}| = 417, \quad |\mathsf{Acc}^{4,m}| = 138, \quad |\mathsf{App}^{4,f}| = 375, \quad |\mathsf{Acc}^{4,f}| = 131$$
$$|\mathsf{App}^{5,m}| = 191, \quad |\mathsf{Acc}^{5,m}| = 53, \quad |\mathsf{App}^{5,f}| = 393, \quad |\mathsf{Acc}^{5,f}| = 94$$
$$|\mathsf{App}^{6,m}| = 373, \quad |\mathsf{Acc}^{6,m}| = 22, \quad |\mathsf{App}^{6,f}| = 341, \quad |\mathsf{Acc}^{6,f}| = 24$$

That App is the disjoint union of $\mathsf{App}^1, \ldots, \mathsf{App}^6$ follows from the previous sentences.

7 Example: Fresh evidence appeal for Leighton Hay's murder conviction

Leighton Hay is one of two men convicted of murdering a man in an Ontario nightclub in 2002. The other man, Gary Eunich, is certainly guilty, but evidence against Hay is weak–much weaker, in my opinion and in the opinion of the Association in Defense of the Wrongly Accused (AIDWYC), than should have been necessary to convict. A good, short summary about the case can be found here: http://www.theglobeandmail.com/news/national/defence-prosecution-split-on-need-for-forensic-hair-testing/article1367543/

The prosecution's case relies strongly on the testimony of one witness, Leisa Maillard, who picked (a 2 year old picture of) Hay out of a photo lineup of 12 black men of similar age, and said she was 80% sure that he was the shooter. There were a number of other witnesses, none of whom identified Hay as one of the killers. Ms. Malard's testimony is weak in a number of ways (e.g. she failed to identify him in a lineup a week after the shooting, and at two trials when she picked out Gary Eunich instead), but here we will be concerned with only one of them: she described the unknown killer as having 2-inch "picky dreads," whereas Hay had short-trimmed hair when he was arrested the morning after the murder. Thus, the police introduced the theory that Hay cut his hair during the night, between the murder and his arrest. In support of the theory, they offered as evidence a balled-up newspaper containing hair clippings that was found at the top of the garbage in the bathroom used by Hay. Their theory, in more detail, is that the known killer, Gary Eunich, cut Hay's hair and beard during the night between the murder and the arrests, using the newspaper to catch the discarded hair, then emptied most of the discarded hair into the toilet; and crucially, a hundred-or-so short hair clippings remained stuck to the newspaper, due perhaps to being lighter than the dreads. It is the origin of those hair clippings that we are primarily concerned

Name in proof	Max width (micrometers)	Count
bin_1	0 to 112.5	10
bin_2	112.5 to 137.5	20
bin_3	137.5 to 162.5	40
bin_4	162.5 to 187.5	19

Table 1: Measurements of 89 hairs found in a balled-up newspaper at the top of Hay's bathroom garbage. Forensic experts on both sides agreed that the hairs in bin_3 and bin_4 are very likely beard hairs, and that the hairs in bin_1 and bin_2 could be either beard or scalp hairs.

Max width (micrometers)	Count
12.5 to 37.5	3
37.5 to 62.5	28
62.5 to 87.5	41
87.5 to 112.5	17
112.5 to 137.5	1

Table 2: Measurements of Hay's scalp hairs obtained at the request of AIDWYC in 2010. Note that the first 4 bins are contained in bin_1 from Table 1. Samples of Hay's beard hairs were not taken and measured in 2010 because the forensic hair experts advised that beard hairs get thicker as a man ages.

with here; Hay has always said that the clippings were from a recent beard-only trim. If that is so, then the newspaper clippings are not at all inculpatory, and knowing this could very well have changed the jury's verdict, since the clippings –as hard as this is to believe– were the main corroborating evidence in support of Ms. Malard's eye witness testimony.

Both sides, defense and prosecution, agree that the newspaper clippings belong to Hay, and that either they originated from his beard and scalp (prosecution's theory), or just his beard (defense's theory). We will try to prove, from reasonable assumptions, that it is more likely that the hair clippings were the product of a beard-only trim than it is that they were the product of a beard and scalp trim.

On 8 Nov 2013 the Supreme Court of Canada granted Hay a new trial in a unanimous decision, based on the new expert analysis of the hair clippings. We do not yet know whether the prosecution will attempt to again use the hair clippings as evidence against Hay.

7.1 High-level argument

In 2002, the prosecution introduced the theory that Hay was the second gunman and must have had his dreads cut off and hair trimmed short during the night following the murder. It is clear that they did this to maintain the credibility of their main witness. In 2012, after the new forensic tests ordered by AIDWYC proved that at least most of the hairs found in Hay's bathroom were (very likely) beard hairs, the prosecution changed their theory to

accommodate, now hypothesizing that the hairs came from the combination of beard and scalp trims with the same electric razor, using the newspaper to catch the clipped hairs for both trims. Intuitively, that progression of theories is highly suspicious.

On the other hand, perhaps the hairs *did* come from the combination of a beard and scalp trim, and the prosecution was simply careless in formulating their original theory. We cannot dismiss the newspaper hairs evidence just because we do not respect the reasoning and rhetoric employed by the prosecution. The argument below takes the prosecution's latest theory seriously. At a high level, the argument has the following structure:

1. There are *many* distinct theories of how the hypothesized beard and scalp trims could have happened. In the argument below, we introduce a family of such theories indexed by the parameters $\alpha_{\min}$ and $\alpha_{\max}$.
2. Most of the theories in that family are bad for the prosecution; they result in a model that predicts the data worse than the defense's beard-trim-only theory.
3. The prosecution cannot justify choosing from among just the theories that are good for them, or giving such theories greater weight.

We will deduce how the parameters $\alpha_{\min}$ and $\alpha_{\max}$ must be set in order for the prosecution's theory to have predictive power as good as the defense's theory, and we will find that the parameters would need to be set to values that have no reasonable justification (without referring to the measurements). If the assumptions from which we derive the parametric theory are reasonable (e.g. the fixed prior over distributions for Hay's beard hair widths, and the fixed distribution for Hay's scalp hair widths), then we can conclude that the newspaper hair evidence is not inculpatory.

Though the argument to follow is unquestionably an example of Bayesian analysis, I prefer to use the language of frequencies and repeatable events rather than degrees of belief. One could just as well use the language of degrees of belief, with no changes to the axioms.

We posit constraints on a randomized simulation model of the crime and evidence, which is applicable not just to Hay's case, but also to a number of very-similar hypothetical cases (in some of which the suspect is guilty) taken from an implicitly-constrained distribution D. The probabilities are just parameters of the model, and in principle we judge models according to how often they make the correct prediction when a case is chosen at random from D. In the argument below, we don't use D directly, but rather use a distribution over a small number of random variables that are meaningful in D, namely the joint distribution for the random variables:

$$\mathsf{G}, \mathsf{Clipped}, \mathsf{Mix}, \mathsf{BParams}, \mathsf{H}, \mathsf{Widths}$$

Some of the most significant assumptions for the argument:

1. The prior chosen for the suspect's beard hair-width distribution is fair and reasonable.[33] This is Simplifying Assumption 3. It is probably the most objectionable of the assumptions. I give some criticisms of it in Section 7.3.

[33]The reason we use a prior for the suspect's beard hair width distribution is that Leighton Hay's beard hair widths were never sampled; that decision was on the advice of one of the hair forensics experts, who said that a man's beard hairs tend to get thicker as he ages.

2. The distribution for the suspect's scalp hair widths, based on the samples taken in 2010, is fair and reasonable (Simplifying Assumption 5).
3. The simulation model, on runs where the suspect is guilty (and thus the newspaper hair evidence comes from a combined beard and scalp trim), chooses uniformly at random (Simplifying Assumption 2) from a sufficiently large range the ratio

$$\frac{\text{P(random clipped hair came from beard, given only that it ended up in the newspaper)}}{\text{P(random clipped hair came from the scalp, given only that it ended up in the newspaper)}}$$

   Specifically that range is $[\frac{\alpha_{\min}}{1-\alpha_{\min}}, \frac{\alpha_{\max}}{1-\alpha_{\max}}]$. The axioms enforce no constraints about $\alpha_{\min}$ and $\alpha_{\max}$ except for $0 < \alpha_{\min} < \alpha_{\max} < 1$, but the hypotheses of Claims 5 and 6 assert significant constraints; it turns out that in order for the likelihood ratio to be $\geqslant 1$, the prosecution needs to make an extreme assumption about $\alpha_{\min}$ and $\alpha_{\max}$. Intuitively, assuming the suspect is guilty, both prosecution and defense are still very ignorant (before seeing the newspaper hair measurements) of how exactly the suspect trimmed his beard and scalp, e.g. in what order, how exactly he used the newspaper, and how exactly he emptied most of the clippings into the toilet, all of which would influence the above ratio. The hypotheses of Claims 5 and 6 formalize that intuition in different ways, which are close to equivalent, but nonetheless I think Claim 6 is significantly easier to understand and accept.
4. The suspect in the simulation model does not have an unusually low ratio of scalp hairs to beard hairs. This is Assumption 16. We can improve the current argument, if we wish, by having the simulation model choose that ratio from some prior distribution, and doing so actually makes results in a version of Claim 6 that is *better* for the defense.

7.2 Argument

A completely-formal version of this argument, which strictly adheres to the definition of *vaguely-interpreted formal proof*, will be included in my thesis. That includes explicit types for each symbol, with each type labeled as an assumption or simplifying assumption. The formalization is mostly straight-forward; the only part that requires some thought is the formalization of random variables and the P(proposition | proposition) syntax.
I will often use the following basic facts. In the completely-formal proof they would be axioms in Γ_{assum} that use only symbols in $\mathcal{L}_{\mathsf{math}}$, and thus should be accepted by any member in the intended audience of the proof.

- For t_1, t_2, t_3 boolean-valued terms:

$$\mathrm{P}(t_1, t_2 \mid t_3) = \mathrm{P}(t_1 \mid t_2, t_3)\mathrm{P}(t_2 \mid t_3)$$

- For X a continuous random variable with conditional density function d_X whose domain S is a polygonal subset of $\mathbb{R}^n$ for some n:

$$\mathrm{P}(t_1 \mid t_2) = \int_{x \in S} \mathrm{P}(t_1 \mid t_2, X = x)\, d_X(x \mid t_2)$$

$\mathbf{bin}_1, \mathbf{bin}_2, \mathbf{bin}_3, \mathbf{bin}_4$ are constants denoting the four micrometer-intervals from Table 1. Formally, they belong to their own sort, which has exactly 4 elements in every model. We do not actually have micrometer intervals in the ontology of the proof, so we could just as well use $\{1, 2, 3, 4\}$, but I think that would be confusing later on. **Bins** is the sort $\{\text{bin}_1, \text{bin}_2, \text{bin}_3, \text{bin}_4\}$.
Throughout this writeup, $\vec{b} = b_1, \ldots, b_{89}$ is a fixed ordering of the newspaper hair measurements shown in Table 1. Specifically, each b_i is one of the constants $\text{bin}_1, \text{bin}_2, \text{bin}_3$, or bin_4; bin_1 appears 10 times, bin_2 20 times, bin_3 40 times, and bin_4 19 times.

$\vec{\mathbf{p}}$ abbreviates $\langle p_1, p_2, p_3 \rangle$.
$\mathbf{p}_4$ abbreviates $1 - p_1 - p_2 - p_3$ (except in Claim 8, as noted there also).

G is the boolean simulation random variable that determines if the suspect in the current run is guilty. I write just G to abbreviate $\mathsf{G} = \text{true}$ and $\overline{\mathsf{G}}$ to abbreviate $\mathsf{G} = \text{false}$.

Clipped is a simulation random variable whose value is determined by G. When G is false, Clipped is the set of beard hair fragments that fall from the suspect's face when he does a full beard trim with an electric trimmer[34] several days before the murder took place. When G is true, Clipped is the set of beard and scalp hair fragments that fall from the suspect's head when he does a full beard trim and a full scalp trim (the latter after cutting off his two-inch dreds) with the same electric trimmer. This includes any such fragments that were flushed down the sink or toilet, but not including –in the case that the suspect is guilty– hair fragments that were part of his 2-inch "picky dreads."

H is a simulation random variable whose distribution is the uniform distribution over Clipped, i.e. it is a random hair clipping.

BParams is the simulation random variable that gives the parameters of the suspect's **b**eard hair width distribution.

Mix is the simulation random variable that gives the the mixture parameter that determine's the prosecution's newspaper hair width distribution given the beard and scalp hair width distributions.

NOTATION: BParams and Mix will usually be hidden in order to de-clutter equations and to fit within the page width. Wherever you see $\vec{p}$ or $\langle p_1, p_2, p_3 \rangle$ where a boolean-valued term is expected, that is an abbreviation for $\mathsf{BParams} = \vec{p}$ or $\mathsf{BParams} = \langle p_1, p_2, p_3 \rangle$, respectively. Similarly, I write just α as an abbreviation for $\mathsf{Mix} = \alpha$.

B is the set from which our prior for the suspect's beard hair width distribution is defined. It

[34]The police collected an electric trimmer that was found, unhidden, in Hay's bedside drawer, which Hay has always said he used for trimming his beard.

is the set of triples $\langle p_1, p_2, p_3 \rangle \in [0,1]^3$ such that $p_1 \leqslant p_2, p_3, p_4$ and $\langle p_1, p_2, p_3, p_4 \rangle$ is unimodal when interpreted as a discrete distribution where p_i is the probability that the width of a hair randomly chosen from the suspect's scalp (in 2002) falls in bin i.

$\mathrm{P}(t_1 \mid t_2)$ is the notation we use for the Bayesian/simulation distribution over the random variables $\mathsf{G}, \mathsf{Clipped}, \mathsf{Mix}, \mathsf{BParams}, \mathsf{H}, \mathsf{Widths}$, where t_1 and t_2 are terms taking on boolean values; it is the probability over runs of the simulation that t_1 evaluates to true given that t_2 evaluates to true.

Widths is the simulation random variable that gives the approximate widths (in terms of the 4 intervals bin_j) of the 89 hair clippings that end up in the balled-up newspaper.

When the variables $\vec{p}$ and α appear unbound in an axiom, I mean for them to be implicitly quantified in the outermost position like so: $\forall \vec{p} \in \mathrm{B}$ and $\forall \alpha \in [\alpha_{\min}, \alpha_{\max}]$.

When X is a continuous random variable with a density function, d_X denotes that function.

Definition 4. We are aiming to show that from reasonable assumptions, the following likelihood ratio is less than 1, meaning that the defense's theory explains the newspaper hairs evidence at least as well as the prosecution's theory.

$$\textit{likelihood-ratio} \coloneqq \frac{\mathrm{P}(\mathsf{Widths} = \vec{b} \mid \mathsf{G})}{\mathrm{P}(\mathsf{Widths} = \vec{b} \mid \overline{\mathsf{G}})}$$

Assumption 9. The values of $\mathsf{BParams}$ and Mix are chosen independently of each other and G (whether or not the suspect is guilty). Hence the defense and prosecution have the same prior for the suspect's beard hair width distribution.
For $t \in \{\text{true}, \text{false}\}$:

$$d_{\langle \mathsf{BParams}, \mathsf{Mix} \rangle}(\vec{p}, \alpha \mid \mathsf{G} = t) = d_{\mathsf{BParams}}(\vec{p}) \cdot d_{\mathsf{Mix}}(\alpha)$$

$\alpha_{\min}$ and $\alpha_{\max}$ are constants in $(0,1)$ such that $\alpha_{\min} < \alpha_{\max}$.

Simplifying Assumption 2. The prior distribution for the mixture parameter Mix is the uniform distribution over $[\alpha_{\min}, \alpha_{\max}]$.

$$d_{\mathsf{Mix}}(\alpha) = \begin{cases} 1/(\alpha_{\max} - \alpha_{\min}) & \text{if } \alpha \in [\alpha_{\min}, \alpha_{\max}] \\ 0 & \text{otherwise} \end{cases}$$

Simplifying Assumption 3. The prior distribution for the parameters of the suspect's beard hair width distribution is the uniform distribution over the set $\mathrm{B} \subseteq [0,1]^3$ defined above.

$$d_{\mathsf{BParams}}(\vec{p}) = \begin{cases} 1/\|\mathrm{B}\| & \text{if } \vec{p} \in \mathrm{B} \\ 0 & \text{otherwise} \end{cases}$$

News$(h) =$ true iff the hair clipping h ends up in the balled-up newspaper.
Beard$(h) =$ true (respectively **Scalp**$(h) =$ true) iff hair clipping h came from the suspect's beard (respectively scalp).

Assumption 10. Both prosecution and defense agreed that all the hairs in the newspaper came from the suspect's beard or scalp, and not both.[35]

$$\text{Scalp}(h) = \neg\text{Beard}(h)$$

width is the function from $\mathsf{Clipped}$ to $\{\text{bin}_1, \text{bin}_2, \text{bin}_3, \text{bin}_4\}$ such that width(h) is the interval in which the maximum-width of hair clipping h falls.

Simplifying Assumption 4. In the simulation model, the hairs that ended up in the newspaper are chosen independently at random with replacement from some hair-width distributions.

$$\text{P}(\mathsf{Widths} = \vec{b} \mid \mathsf{G}, \vec{p}, \alpha) = \prod_{i=1}^{89} \text{P}(\text{width}(\mathsf{H}) = b_i \mid \text{News}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha)$$

$$\text{P}(\mathsf{Widths} = \vec{b} \mid \overline{\mathsf{G}}, \vec{p}) = \prod_{i=1}^{89} \text{P}(\text{width}(\mathsf{H}) = b_i \mid \text{News}(\mathsf{H}), \overline{\mathsf{G}}, \vec{p})$$

Claim 4. We can write the width distribution of newspaper hairs in terms of the width distributions of beard and scalp hairs, together with the probability that a random newspaper hair is a beard hair.

$$\begin{aligned} & \text{P}(\text{width}(\mathsf{H}) = b_i \mid \text{News}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha) \\ = \; & \text{P}(\text{width}(\mathsf{H}) = b_i \mid \text{Beard}(\mathsf{H}), \text{News}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha)\ \text{P}(\text{Beard}(\mathsf{H}) \mid \text{News}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha) \\ + \; & \text{P}(\text{width}(\mathsf{H}) = b_i \mid \text{Scalp}(\mathsf{H}), \text{News}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha)\ \text{P}(\text{Scalp}(\mathsf{H}) \mid \text{News}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha) \end{aligned}$$

Proof. Follows from Assumption 10. □

Assumption 11. In the defense's model (not guilty $\overline{\mathsf{G}}$), all the newspaper hair came from a beard trim, and so the mixture parameter is irrelevant.

$$\begin{aligned} & \text{P}(\text{width}(\mathsf{H}) = b_i \mid \text{News}(\mathsf{H}), \overline{\mathsf{G}}, \vec{p}, \alpha) \\ = \; & \text{P}(\text{width}(\mathsf{H}) = b_i \mid \text{Beard}(\mathsf{H}), \text{News}(\mathsf{H}), \overline{\mathsf{G}}, \vec{p}) \end{aligned}$$

Assumption 12. Given that a clipped hair came from the suspect's beard, the hair's width is independent of whether the suspect is guilty in this run of the simulation. Thus the defense and prosecution models use the same distribution of hair widths for the suspect's beard.

$$\begin{aligned} & \text{P}(\text{width}(\mathsf{H}) = b_i \mid \text{Beard}(\mathsf{H}), \text{News}(\mathsf{H}), \overline{\mathsf{G}}, \alpha, \vec{p}) \\ = \; & \text{P}(\text{width}(\mathsf{H}) = b_i \mid \text{Beard}(\mathsf{H}), \text{News}(\mathsf{H}), \mathsf{G}, \alpha, \vec{p}) \\ = \; & \text{P}(\text{width}(\mathsf{H}) = b_i \mid \text{Beard}(\mathsf{H}), \text{News}(\mathsf{H}), \alpha, \vec{p}) \end{aligned}$$

[35] "Not both" actually ignores the issue of sideburn hairs, whose widths can be intermediate between scalp and beard hair widths. Doing this is favourable for the prosecution.

Assumption 13. We finally give the precise meaning of the simulation's mixture parameter random variable Mix. It is the probability, when the suspect is guilty, that a randomly chosen hair clipping came from the suspects beard *given* that it ended up in the newspaper.

$$\alpha = \mathrm{P}(\mathrm{Beard}(\mathsf{H}) \mid \mathrm{News}(\mathsf{H}), \mathsf{G}, \vec{p}, \mathsf{Mix}{=}\alpha)$$

$$1-\alpha = \mathrm{P}(\mathrm{Scalp}(\mathsf{H}) \mid \mathrm{News}(\mathsf{H}), \mathsf{G}, \vec{p}, \mathsf{Mix}{=}\alpha)$$

Assumption 14. The precise meaning of the simulation random variable $\mathsf{BParams}$. Recall that p_4 abbreviates $1-p_1-p_2-p_3$. For $j \in \{1,2,3,4\}$:

$$p_j = \mathrm{P}(\mathrm{width}(\mathsf{H}){=}\mathrm{bin}_j \mid \mathrm{Beard}(\mathsf{H}), \mathsf{BParams}{=}\langle p_1, p_2, p_3\rangle, \mathrm{News}(\mathsf{H}))$$

Simplifying Assumption 5. We use a completely-fixed distribution for the suspect's scalp hair, namely the one that maximizes the probability of obtaining the hair sample measurements from Table 2 when 90 hairs are chosen independently and uniformly at random from the suspect's scalp.

$$\mathrm{P}(\mathrm{width}(\mathsf{H}) = b_i \mid \mathrm{Scalp}(\mathsf{H}), \mathsf{G}, \alpha, \vec{p}) = \begin{cases} {}^{89}\!/_{90} & \text{if } i = 1 \\ {}^{1}\!/_{90} & \text{if } i = 2 \\ 0 & \text{if } i = 3,4 \end{cases}$$

The next axiom and claim give the main result, and the later Claim 6 is (almost) a corollary of Claim 5.

Assumption 15. If $\frac{\mathrm{P}(\mathsf{Widths}=\vec{b}\mid\mathsf{G})}{\mathrm{P}(\mathsf{Widths}=\vec{b}\mid\overline{\mathsf{G}})} \leqslant 1$ (i.e. *likelihood-ratio* $\leqslant 1$), then $\langle$the newspaper hair evidence is neutral or exculpatory$\rangle$.[36]

Claim 5. If $\alpha_{\min} \leqslant .849$ then $\frac{\mathrm{P}(\mathsf{Widths}=\vec{b}\mid\mathsf{G})}{\mathrm{P}(\mathsf{Widths}=\vec{b}\mid\overline{\mathsf{G}})} < 1$

The proof of Claim 5 is outlined formally below, after Claim 6.

With the introduction of a new parameter and a mild assumption about its values (Assumption 16, the ratio on the left side being the new parameter), we will obtain a corollary of Claim 5 that is easier to interpret.

We do not know what the ratio of beard to scalp hairs on Hay's head was on the date of the murder, and it is not hard to see that a higher value of $\mathrm{P}(\mathrm{Beard}(\mathsf{H}) \mid \mathsf{G}, \vec{p}, \alpha)$ is favourable for the prosecution.[37] We do, however, know that the unknown shooter's beard was described as "scraggly" and "patchy" by eye witnesses, and we have no reason to think that LH had a smaller than average number of scalp hairs. Thus it is a conservative approximation (from the perspective of the prosecution) to assume that Hay had a great quantity of beard hairs for

[36] The text in brackets is a constant predicate symbol.

[37] Raising the value makes both models worse, but it hurts the prosecution's model less since the prosecution's model can accommodate by lowering $\alpha_{\min}$ and $\alpha_{\max}$.

a man (40,000), and an average quantity of scalp hairs for a man with black hair (110,000).[38] Thus we assume:

Assumption 16.

$$\frac{\mathrm{P}(\mathrm{Beard}(\mathsf{H}) \mid \mathsf{G}, \vec{p}, \alpha)}{\mathrm{P}(\mathrm{Scalp}(\mathsf{H}) \mid \mathsf{G}, \vec{p}, \alpha)} \leqslant 4/11$$

Claim 6. The hypothesis of Assumption 15 also follows if we assume Assumption 16 and that *the uniform prior over* Mix *gives positive density to a model where a random clipped beard hair is* $\leqslant 15$ *times more likely to end up in the newspaper as a random clipped scalp hair*:
If there exists $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ and $\vec{p} \in \mathrm{B}$ such that

$$\frac{\mathrm{P}(\mathrm{News}(\mathsf{H}) \mid \mathrm{Beard}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha)}{\mathrm{P}(\mathrm{News}(\mathsf{H}) \mid \mathrm{Scalp}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha)} \leqslant 15$$

then

$$\frac{\mathrm{P}(\mathsf{Widths} = \vec{b} \mid \mathsf{G})}{\mathrm{P}(\mathsf{Widths} = \vec{b} \mid \overline{\mathsf{G}})} < 1$$

Proof. Let $\alpha, \vec{p}$ be as in the hypothesis.
From basic rules about conditional probabilities:

$$\frac{\alpha}{1-\alpha} = \frac{\mathrm{P}(\mathrm{Beard}(\mathsf{H}) \mid \mathrm{News}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha)}{\mathrm{P}(\mathrm{Scalp}(\mathsf{H}) \mid \mathrm{News}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha)} = \frac{\mathrm{P}(\mathrm{News}(\mathsf{H}) \mid \mathrm{Beard}(\mathsf{H})\mathsf{G}, \vec{p}, \alpha)\ \mathrm{P}(\mathrm{Beard}(\mathsf{H}) \mid \mathsf{G}, \vec{p}, \alpha)}{\mathrm{P}(\mathrm{News}(\mathsf{H}) \mid \mathrm{Scalp}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha)\ \mathrm{P}(\mathrm{Scalp}(\mathsf{H}) \mid \mathsf{G}, \vec{p}, \alpha)} \tag{1}$$

Using the inequality from the hypothesis and Assumption 16, solve for α in (1). This gives $\alpha \leqslant 0.84507$. Since $\alpha_{\min} \leqslant \alpha$ we have $\alpha_{\min} \leqslant .84507$, so we can use Claim 5 to conclude that the likelihood ratio is less than 1. □

Simplifying Assumption 6 (hypothesis of Claim 6)**.** There exists $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ and $\vec{p} \in \mathrm{B}$ such that

$$\frac{\mathrm{P}(\mathrm{News}(\mathsf{H}) \mid \mathrm{Beard}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha)}{\mathrm{P}(\mathrm{News}(\mathsf{H}) \mid \mathrm{Scalp}(\mathsf{H}), \mathsf{G}, \vec{p}, \alpha)} \leqslant 15$$

Goal Sentence 1. ⟨the newspaper hair evidence is neutral or exculpatory⟩

Proof. From Simplifying Assumption 6, Claim 6, and Assumption 15. □

[38] Trustworthy sources for these numbers are hard to find. 40,000 is just the largest figure I found amongst untrustworthy sources, and 110,000 is a figure that appears in a number of untrustworthy sources. If this troubles you, consider the ratio a parameter whose upper bound we can argue about later.

Proof of Claim 5

Note: there is nothing very interesting about this proof; it is basically just a guide for computing the likelihood-ratio as a function of $\alpha_{min}, \alpha_{max}$.

To compute the integrals, I will break up the polygonal region B into several pieces which are easier to handle with normal Riemann integration over real intervals.
Let B_1 be the subset of B where $p_2 > p_3 \geqslant p_4$
B_2 the subset of B where $p_3 > p_2 > p_4$
B_3 the subset of B where $p_3 > p_4 \geqslant p_2$
B_4 the subset of B where $p_4 > p_3 \geqslant p_2$

Claim 7. B is the disjoint union of $\mathrm{B}_1, \mathrm{B}_2, \mathrm{B}_3, \mathrm{B}_4$.

Claim 8. In the scope of this claim, p_4 is a normal variable, not an abbreviation for $1 - p_1 - p_2 - p_3$.

$$\int\limits_{\vec{p}=\langle p_1,p_2,p_3\rangle \in \mathrm{B}_1} t(p_1,p_2,p_3,1-p_1-p_2-p_3)d\vec{p} = \int\limits_{p_1=0}^{1/4} \int\limits_{p_4=p_1}^{\frac{1-p_1}{3}} \int\limits_{p_3=p_4}^{\frac{1-p_1-p_4}{2}} t(p_1,1-p_1-p_3-p_4,p_3,p_4')dp_1dp_4dp_3$$

$$\int\limits_{\vec{p}=\langle p_1,p_2,p_3\rangle \in \mathrm{B}_2} t(p_1,p_2,p_3,1-p_1-p_2-p_3)d\vec{p} = \int\limits_{p_1=0}^{1/4} \int\limits_{p_4=p_1}^{\frac{1-p_1}{3}} \int\limits_{p_2=p_4}^{\frac{1-p_1-p_4}{2}} t(p_1,p_2,1-p_1-p_2-p_4,p_4)dp_1dp_4dp_2$$

$$\int\limits_{\vec{p}=\langle p_1,p_2,p_3\rangle \in \mathrm{B}_3} t(p_1,p_2,p_3,1-p_1-p_2-p_3)d\vec{p} = \int\limits_{p_1=0}^{1/4} \int\limits_{p_2=p_1}^{\frac{1-p_1}{3}} \int\limits_{p_4=p_2}^{\frac{1-p_1-p_2}{2}} t(p_1,p_2,1-p_1-p_2-p_4,p_4)dp_1dp_2dp_4$$

$$\int\limits_{\vec{p}=\langle p_1,p_2,p_3\rangle \in \mathrm{B}_4} t(p_1,p_2,p_3,1-p_1-p_2-p_3)d\vec{p} = \int\limits_{p_1=0}^{1/4} \int\limits_{p_2=p_1}^{\frac{1-p_1}{3}} \int\limits_{p_3=p_2}^{\frac{1-p_1-p_2}{2}} t(p_1,p_2,p_3,1-p_1-p_2-p_3)dp_1dp_2dp_3$$

Claim 9. $\|\mathrm{B}\| = 1/36$

Proof. The measure of B_j can be computed by standard means by substituting 1 in for $t(\ldots)$ in the right side of the j-th equation of Claim 8. We find that $\|\mathrm{B}_1\| = \|\mathrm{B}_2\| = \|\mathrm{B}_3\| = \|\mathrm{B}_4\| = 1/144$. Hence $\|\mathrm{B}\| = 1/36$ follows from Claim 7. □

Claim 10. Simplified forms amenable to efficient computation:

$$\mathrm{P}(\mathsf{Widths}=\vec{b} \mid \overline{\mathsf{G}}, \langle p_1,p_2,p_3\rangle) = p_1^{10}p_2^{20}p_3^{40}p_4^{19}$$

$$\mathrm{P}(\mathsf{Widths}=\vec{b} \mid \mathsf{G}, \langle p_1,p_2,p_3\rangle, \alpha) = (p_1\alpha + {}^{89}\!/_{90}(1-\alpha))^{10}(p_2\alpha + {}^{1}\!/_{90}(1-\alpha))^{20}(p_3\alpha)^{40}(p_4\alpha)^{19}$$

Proof. The first equation follows easily from Simplifying Assumption 4 and Assumption 14. The second follows easily from Simplifying Assumption 4, Axioms 14 and 13, and Claim 4. □

From the next fact and Claim 8 we can compute the two terms of the likelihood ratio for fixed $\alpha_{\min}$ and $\alpha_{\max}$.

Claim 11.

$$\mathrm{P}(\mathsf{Widths} = \vec{b} \mid \mathsf{G}) = \int_{\alpha \in [\alpha_{\min}, \alpha_{\max}]} \int_{\vec{p} \in \mathrm{B}} \mathrm{P}(\mathsf{Widths} = \vec{b} \mid \mathsf{G}, \vec{p}, \alpha)\, d_{\langle \mathsf{BParams}, \mathsf{Mix} \rangle}(\vec{p}, \alpha \mid \mathsf{G})$$

$$= \frac{1}{(\alpha_{\max} - \alpha_{\min})\|\mathrm{B}\|} \sum_{i \in \{1,2,3,4\}} \int_{\alpha \in [\alpha_{\min}, \alpha_{\max}]} \int_{\vec{p} \in \mathrm{B}_i} \mathrm{P}(\mathsf{Widths} = \vec{b} \mid \mathsf{G}, \vec{p}, \alpha)$$

$$\mathrm{P}(\mathsf{Widths} = \vec{b} \mid \overline{\mathsf{G}}) = \int_{\vec{p} \in \mathrm{B}} \mathrm{P}(\mathsf{Widths} = \vec{b} \mid \overline{\mathsf{G}}, \vec{p})\, d_{\mathsf{BParams}}(\vec{p} \mid \overline{\mathsf{G}})$$

$$= \frac{1}{\|\mathrm{B}\|} \sum_{i \in \{1,2,3,4\}} \int_{\vec{p} \in \mathrm{B}_i} \mathrm{P}(\mathsf{Widths} = \vec{b} \mid \overline{\mathsf{G}}, \vec{p})$$

Proof. The first equation follows just from $\vec{p}, \alpha \mapsto \mathrm{P}(\mathsf{Widths} = \vec{b} \mid \mathsf{G}, \vec{p}, \alpha)$ being an integrable function and $d_{\langle \mathsf{BParams}, \mathsf{Mix} \rangle}(\vec{p}, \alpha \mid \mathsf{G})$ being the conditional density function for $\langle \mathsf{Mix}, \mathsf{BParams} \rangle$ given $\mathsf{G} = \text{true}$.
The second equation follows from Claim 7, Simplifying Assumptions 2 and 3, and the fact that $\vec{p}, \alpha \mapsto \mathrm{P}(\mathsf{Widths} = \vec{b} \mid \mathsf{G}, \vec{p}, \alpha)$ is bounded. The first and fourth of those facts suffice to show that the integral over B is equal to the sum of the integrals over the sets B_j.
Justifications for the third and fourth equations are similar to those for the first and second. □

As of now I've mostly used Mathematica's numeric integration, which doesn't provide error bounds, to evaluate the intervals, but there are also software packages one can use that provide error bounds.
The likelihood ratio achieves its maximum of ≈ 1.27 when $\alpha_{\min}$ and $\alpha_{\max}$ are practically equal (unsurprising, as that allows the prosecution model to choose the best mixture parameter) and around .935; Plot 7.2 illustrates this, showing the likelihood ratio as a function of $\alpha_{\min}$ when $\alpha_{\max} - \alpha_{\min} = 10^{-6}$. To prove Claim 5 we need to look at parameterizations of $\alpha_{\min}, \alpha_{\max}$ similar to the one depicted in Plot 7.2, which shows the likelihood ratio as a function of $\alpha_{\max}$ when $\alpha_{\min} = .849$ (the extreme point in the hypothesis of Claim 5), in which case the likelihood ratio is maximized at $\approx .996$ when $\alpha_{\max} = 1$. In general, for smaller fixed $\alpha_{\min}$, the quantity

$$\max_{\alpha_{\max} \in (\alpha_{\min}, 1)} (\textit{likelihood-ratio}(\alpha_{\min}, \alpha_{\max}))$$

decreases as $\alpha_{\min}$ does. More precisely, Claim 5 follows from the following three propositions in Claim 12. The first has been tested using Mathematica's numerical integration; if it is false, it is unlikely to be false by a wide margin (i.e. taking a value slightly smaller than .849 should suffice). The remaining two have also not been proved, but one can gain good confidence in them by testing plots similar to Figure 7.2 for values of $\alpha_{\min} < .849$. Proving or disproving Claim 12 is just a matter of spending more time on it (or enlisting the help of an expert to do it quickly). But we will see in the next section that the argument is more-vulnerable to attack in other ways.

Claim 12.

1. *likelihood-ratio*$(.849, 1) < .997$
2. For $\alpha_1 < .849$ have *likelihood-ratio*$(\alpha_1, 1) <$ *likelihood-ratio*$(.849, 1)$
3. For $\alpha_1 < .849$ and $\alpha_1 < \alpha_2 < 1$ have *likelihood-ratio*$(\alpha_1, \alpha_2) <$ *likelihood-ratio*$(\alpha_1, 1)$

Figure 1: Likelihood ratio as a function of $\alpha_{\min}$ when $\alpha_{\max} - \alpha_{\min} = 10^{-6}$, obtained by numerical integration.

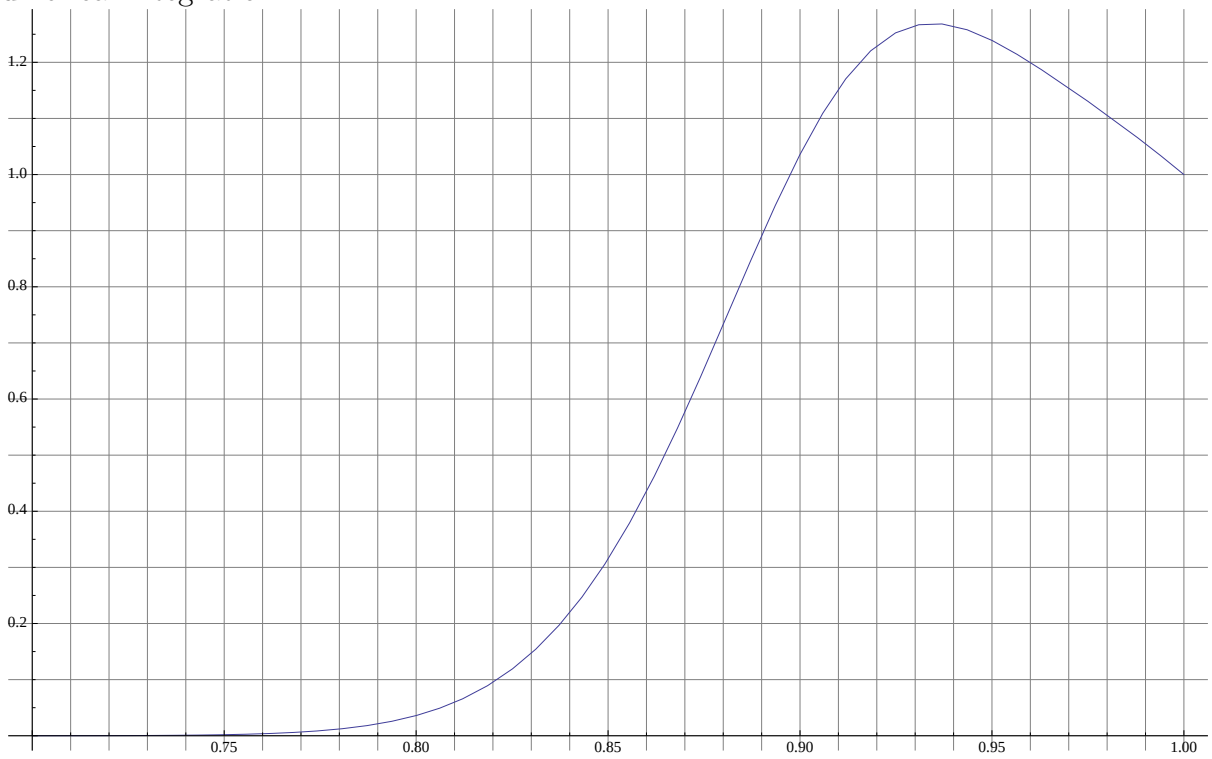
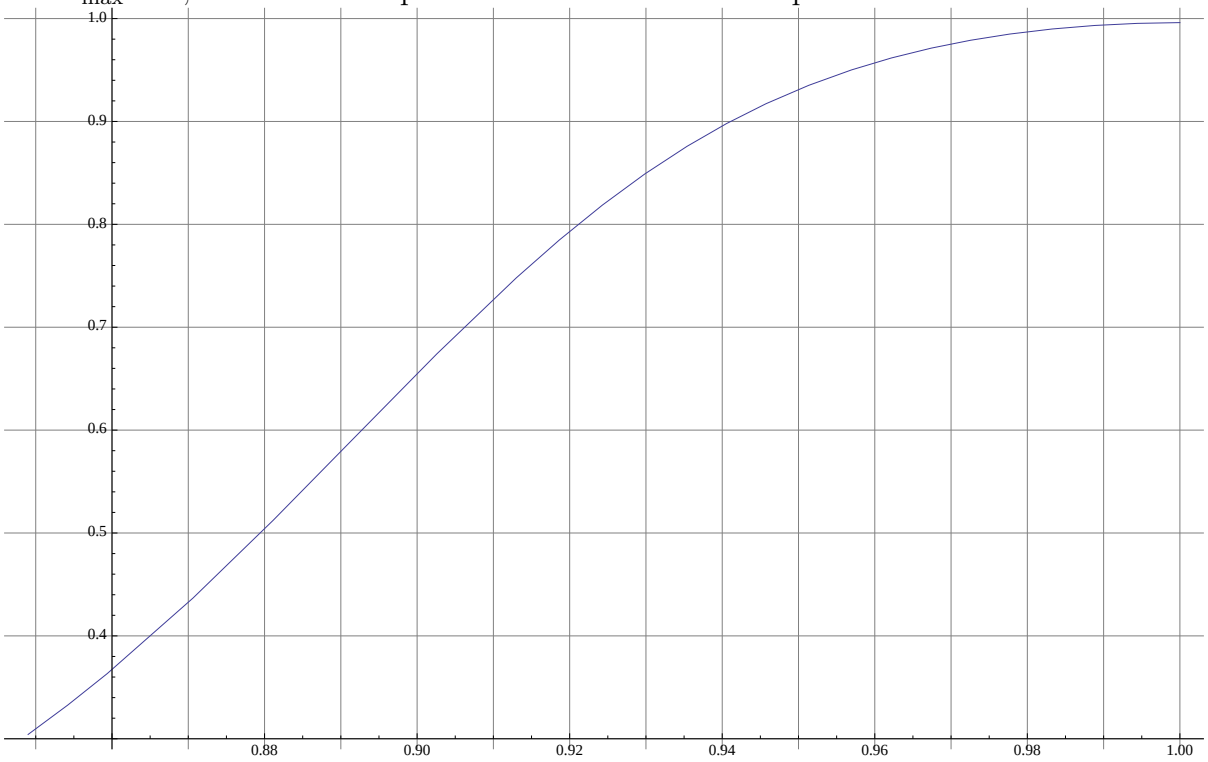


Figure 2: Likelihood ratio as a function of $\alpha_{\max}$ when $\alpha_{\min} = .849$, obtained by numerical integration. The shape of this plot is similar for smaller values of $\alpha_{\min}$, being maximized when $\alpha_{\max} = 1$, which is what parts 2 and 3 of Claim 12 express.



7.3 Criticism of argument

7.3.1 Criticism 1

It is arguable that the prior for the suspect's beard hair width distribution is slightly biased in favor of the defense, in which case the prosecution could **reject Simplifying Assumption 3**. In particular, the average value of the component of $\mathsf{BParams}$ for bin_1, the bin corresponding to the thinnest hairs, is 0.0625.[39] It is best for the defense when the value of that component is 11/89, and best for the prosecution when it is 0, so the prosecution could reasonably insist that a prior is not fair unless the average is at most the mean of those two extremes, which is ≈ 0.0618.

We can raise this criticism in a disciplined way, for example by suggesting an axiom that expresses the above; if x is the value of p_1 that maximizes the probability of the evidence given G = true, and y is the value of the p_1 that maximizes the probability of the evidence given G = false, then $\int_{\vec{p}\in \mathrm{B}} p_1 \leqslant (x+y)/2$.

The defense can respond to the criticism, and I will show that. This requires slightly strengthening the hypotheses of Claims 5 and 6.

[39] Compute by substituting p_1 in for t in each of the four equations of Claim 8, and sum the results.

7.3.2 Criticism 2

The prior for BParams is unreasonable, with respect to measurements of beard hair widths of black men in the literature, in that it never yields a beard hair width distribution that has hairs of width greater than 187.5 micrometers. In terms of the argument, we should reject the (implicit) axioms that constitute the types of width (and/or Widths); according to the semantics of those symbols, their types assert that all the hairs in Leighton Hay's beard and scalp had thickness at most 187.5 micrometers, which is unjustified. Formally, one way to do this would be to suggest new definitions of Bins, width, and Widths. We can do this by suggesting new axioms (some of which are type constraints). Most importantly we should suggest redefining the sort Bins as $\{\text{bin}_1, \ldots, \text{bin}_5\}$, where bin_5 is a new constant. The results of that approach are discussed in Section 7.3.3.

7.3.3 Response to criticisms

We can address both criticisms at once; if we introduce a fifth component of BParams corresponding to the interval $(187.5, \infty)$, and like the first component (probability width is in bin_1) of BParams constrain it to be less than the middle three components (for $\text{bin}_2, \text{bin}_3, \text{bin}_4$), then the average value of the bin_1 component of BParams goes down to $< .057$. We then need to slightly strengthen the hypotheses of the two main claims, changing the parameter .85 in Claim 5 to .835 and the parameter 15 in Claim 6 to 13.9.

7.3.4 An open problem

Though I do not have such a criticism in mind, the prosecution could potentially argue that the prior for Hay's beard hair distribution is still biased, in the sense that it does not take into account everything we know about the beard hair width distributions of young black men or Hay himself, say by referring to literature such as [10] (cited in the documents submitted by expert witnesses from both sides of the trial), or by taking samples of Hay's current beard hair width distribution and somehow adjusting for the increase in width that expert witnesses said is likely, since Hay was only 19 at the time of the murder. Or they could criticize my choice of prior by claiming that it assumes *too much*.[40]

Given that, an ideal proof would have the following form. We would first come up with some relation R over priors for 5-bin distributions, such that $R(f)$ expresses as well as possible (given the constraint of having to complete the proof of the following proposition) that f is "fair and reasonable", . Then, we would find the largest constant $\alpha_0 \in (0, 1)$ such that we can prove:

[40] Although I expect that would be a bad idea. For example, I found that if we take the prior to be the completely uniform prior over finite distributions for 5 bins, then the results are significantly worse for the prosecution.

For any $f \in R$, if f is used as the prior for the suspect's beard hair width distribution, and $\alpha_{\min} < \alpha_0$, then *likelihood-ratio* < 1

The same goes for Hay's scalp hair width distribution; it would be better to have a broad set of distributions that an adversary can choose from. At the very least, the argument should accommodate the possibility that Hay's scalp hairs have thinned over time, in which case we would make use of the fact that Hay is *not* balding (male pattern balding makes hair follicles, and the hairs they produce, gradually thinner, until the hair follicle is blocked completely).

8 Ongoing work

My continued work on this project involves, first of all, writing more examples of vaguely-interpreted formal proofs and criticisms of them. Secondly, I am working on a web interface for authoring and criticizing vaguely-interpreted formal proofs; PDF/postscript documents are very poorly suited for the purpose, since the page-oriented structure precludes the possibility of documents that make extensive use of collapsing/expanding sections of text. The web interface has benefits such as being able to view the informal semantics for a symbol just by hovering one's cursor over any occurrence of this symbol; this is important, as arguments about problems in the domain described in Section 2 tend to require a much larger number of irreducible objects than typical proofs in mathematics.

One of my major in-progress examples is an argument in support of an assisted suicide policy for Canada (a larger-scale argument than the one from Section 5, which applies only to one court case), which is something I think the vast majority of people believe is well outside the scope of mathematics/formal logic. And that brings me to the primary message of this paper: such beliefs about the limited applicability of formal logic, though understandable, are wrong, and we should be actively disputing them with examples.

A Informal sketch of a theoretical dialogue system

Vaguely-Interpreted Formal Proofs with Finite Models
The dialogue system is asymmetric, with different roles for the critics and the proponent/author of the argument. The definition of a state of a dialogue includes at least the following:

1. A vaguely-interpreted formal proof π with signature Σ and axioms Γ.
2. A set Γ_{ind} of Σ-sentences that are independent of Γ, and are a minimal set of independent sentences in the following sense: for every $A \in \Gamma_{\mathsf{ind}}$, the independence of A from Γ does not follow from the independence of $\Gamma_{\mathsf{ind}} - A$ from Γ.
3. A subsignature Σ_{fin} of Σ that contains all the symbols of $\mathcal{L}_{\mathsf{vague}}$ and any symbols of $\mathcal{L}_{\mathsf{def}}$ that depend on symbols from $\mathcal{L}_{\mathsf{vague}}$.

Let Γ_{fin} denote the axioms of Γ that are well-typed Σ_{fin}-sentences. *All controversial axioms must be in* Γ_{fin}*, and* Γ_{fin} *must have finite models.*

The role of Γ_{ind} is to allow progress in the form of making vagueness *explicit*. Since vagueness, in the framework of vaguely-interpreted formal proofs, is represented by plurality of intended interpretations, there is not yet any provision to distinguish between *fundamentally vague predicates* and *weak axiomatizations of sharply-definable predicates* (and perhaps the definition of vaguely-interpreted formal proof should be modified to accommodate such a thing). In short, if I have $A \in \Gamma_{\mathsf{ind}}$, that means that there are structures that falsify A and structures that satisfy A that I consider both *intended models* of the axioms.[41]

Since symbols in $\mathcal{L}_{\mathsf{vague}}$ may have types that contain sort symbols from $\mathcal{L}_{\mathsf{math}}$ whose intended interpretation is infinite, and since controversial axioms will often include symbols from $\mathcal{L}_{\mathsf{math}}$, converting an argument into a form that fits this schema would sometimes require introducing new versions of sort and term symbols in $\mathcal{L}_{\mathsf{math}}$ that have finite intended interpretations (surely there must be a better way!). For example, in the Berkeley argument from Section 6, we would introduce a second version of the cardinality function $|\cdot| : \mathcal{A} \to \mathbb{N}$ that has type $\mathcal{A} \to \mathbb{N}'$, where $\mathbb{N}'$ is a new sort symbol whose intended interpretation is the first 4526 natural numbers; enough to give a size to each of the relevant sets of applications. This complication is just one of the reasons why I am calling this a sketch of a *theoretical* dialog system.

In my dissertation, I will give general conditions on vaguely-interpreted formal proofs that enable one to convert a proof to the above schema. Intuitively, this amounts to an argument not making indispensable use of infinite objects. My hypothesis is that for arguments about problems in the domain I am interested in (Section 2), doing so is always possible *in principle*.

Dialogue rules and definition of progress

The rules would consist of:

1. Rules for refactoring proofs, and for adding detail by substituting proved lemmas for axioms.
2. Rules for moves similar to those given at the end of Section 3 for criticizing a vaguely-interpreted formal proof.
3. Rules for moves that extend the language.

It turns out to be easy to ensure progress if the language is not extended indefinitely, essentially by requiring the criticisms and responses to criticisms eliminate some finite models.

[41] E.g. if for some reason we needed to axiomatize a predicate *tall*, and had a sentence expressing "People of height 185cm are tall" in Γ_{fin}, and a sentence expressing "People of height 175cm are tall" in Γ_{ind} (asserting there exists an intended model that says 175cm is tall, and another that says 175cm is not tall), then with obvious additional axioms, we would ensure that tallness is vague under the 175cm mark, and sharply defined over the 185cm mark, while in the (175cm,185cm) range we have the option to later either make the predicate more precise, or extend the range of vagueness upward. A sentence expressing "If people of height x cm are tall then people of height x - .1 cm are tall" would be rejected for addition to Γ_{fin} for contradicting the independence of "People of height 175cm are tall", without us needing to specify a minimum height for tallness.

We can then build on that conclusion by allowing moves that extend the language but simultaneously make progress in another way. The gist is this: we allow a move that extends the language $\mathcal{L}$ to $\mathcal{L}'$ and the axioms Γ to Γ' if the number of finite models of Γ_{fin} that can be extended to models of Γ'_{fin} is smaller than the number of finite models of Γ_{fin}.[42] Hence the overall vagueness in the argument must decrease. General expansions of the language of an argument, which often amounts to expanding the scope of the argument, can be used to block progress indefinitely. On the other hand, we can rarely specify all the relevant concepts before an argument begins. I have no better idea for a fix of this problem that to allow any number of expansions of the language that are accompanied by some mark of progress, and a pre-specified limit on the number of general expansions, with no mark of progress, that either party can do. I am very open to suggestions, but I am also hoping that it is only a theoretical problem that won't have a practical impact on this project.

References

[1] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley. *S* cience, 187(4175):398–404, 1975.

[2] William M. Farmer. A simple type theory with partial functions and subtypes. *A* nnals of Pure and Applied Logic, 64(3):211–240, November 1993.

[3] K. Gödel and S. Feferman. *K* urt Gödel: Collected Works: Volume III: Unpublished Essays and Lectures. Collected Works of Kurt Godel. OUP USA, 1995.

[4] G.W. Leibniz, S. Charlotte, and L. Strickland. *L* eibniz and the Two Sophies: The Philosophical Correspondence. Other voice in early modern Europe: Toronto series. Iter Incorporated, 2011.

[5] G.W. Leibniz and L.E. Loemker. *P* hilosophical Papers and Letters. Number v. 1 in Synthese Historical Library. D. Reidel Publishing Company, 1976.

[6] Judea Pearl. *C* ausality: Models, Reasoning and Inference. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[7] Henry Prakken. An abstract framework for argumentation with structured arguments. *A* rgument & Computation, 1(2):93–124, 2010.

[8] Neven Sesardic. Sudden infant death or murder? a royal confusion about probabilities. *T* he British Journal for the Philosophy of Science, 58(2):299–329, 2007.

[9] Roy Sorensen. Vagueness. In Edward N. Zalta, editor, *T* he Stanford Encyclopedia of Philosophy. Winter 2013 edition, 2013.

[42] Actually, $\mathcal{L}$ and Γ are not necessarily from the previous state of the dialogue. They are from the last of the (bounded number of) states of the dialogue when the language was extended in a way that doesn't ensure progress in any way.

[10] Eva Tolgyesi, DW Coble, FS Fang, and EO Kairinen. A comparative study of beard and scalp hair. *J* Soc Cosmet Chem, 34:361–382, 1983.

[11] D. Walton and E. Krabbe. *C* ommitment in Dialogue: Basic concepts of interpersonal reasoning. State University of New York Press, Albany NY, 1995.

[12] D.N. Walton. *I* nformal Logic: A Pragmatic Approach. Cambridge University Press, 2008.