Predicate Logic for Rigorous Deductive Argumentation on Subjective and Vague Questions

Thesis proposal

Dustin Wehr

March 18, 2014

Contents

1 Introduction

Gottfried Leibniz had a dream, long before the formalization of predicate logic, that some day the rigor 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[1])

Unfortunately he was right about the maturity of mankind, and probably still is. But whether or not it is futile to hope that such a dream might one day materialize, there is still the unanswered question of whether it *ought* to be possible.

What is clear to many logicians, mathematicians, and philosophers –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 my project, is to illustrate this understanding. The second goal, a very difficult and lonely one, is to investigate whether such use of rigorous deduction could ever be cost effective, which involves both making such work easier, and explicating its benefits.

There are thousands and thousands of pages by hundreds of scholars that are tangentially related to this work; papers about vagueness in the abstract,[2] the theoretical foundations of Bayesian reasoning,[3] *abstract dialog systems* [Pra10], 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. My project is of the latter kind; I work on important specific problems, attacking general theoretical problems only as-necessary.

Surprisingly, it is the (normative side of the) field of Informal Logic that is probably most related to my project [Wal08][WK95]. 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 1.1). 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 rigor that

[1] From translation of a letter to Sophia of Hanover [LCS11]

[2] See [Sor13], 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 standard models.*

[3] I recommend [Pea09].

they usually follow in their professional work.[4]

Regarding the title of this thesis: there are three aspects of contentious socially-relevant questions that distinguish them from typical mathematics problems: vagueness, subjectiveness, and uncertainty.

Vagueness is not a problem once you become comfortable with the understanding that definitions need to be precisified gradually as an argument[5] proceeds. With mathematics problems we can *usually* axiomatize structures sufficiently-precisely at the beginning of our reasoning. The commonality between proofs in mathematics and the sort of proofs I do in this thesis is that we only need to axiomatize the structures we are thinking about precisely enough for the proofs to go through. Of course, when I say vagueness is not a problem, I do not mean that vague questions can 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 up to the author. An illustrative example of this can be made with Newcomb's Paradox[6]; it is not hard to give two reasonable formalizations of the problem that yield opposite answers.

Subjectiveness just demands some system of interaction between people on the two sides of an argument. Initially I thought that a dialogue system, with rules designed to enforce progress as much as possible, would be appropriate. I have since backed off that idea somewhat, because (1) my progress so far gives a system that is too complicated to be useful in practice (see Section 3.3), and (2) I'm concerned that, even if a simpler system was devised, asking people to commit to a dialogue is asking too much. So I have shifted to an informal 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 with a new proof. 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 suss out *fundamental* sources of disagreement starting from disagreement on some complex question (which is progress!).

Uncertainty is the most difficult of the three complications. Sparsity of information can easily make it impossible to give an *absolutely*-strong deductive argument for or against a given proposition. But interaction is useful here, too. For example, I give a proof in Section 5 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 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[7] 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.

[4] [Ses07] 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.

[5] A sequence of proofs and critiques, as described in Section 3.2.

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

[7] i.e. that the likelihood ratio is much larger than 1.

1.1 Problem Domain / Scope

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 that the only major impediment to rigorous deductive reasoning about socially relevant issues is basically conventional problem-solving difficulty[8]. Of course, it is a strong impediment. 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 one's options are finished heuristic reasoning or unfinished deductive reasoning, it is probably best to go with the former.

Without contentiousness, there is little motive for employing 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 rigor, 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*[9] *we might reason as safely in metaphysics as in mathematics.*
>
> (Gottfried Leibniz, 1678, Letter to Walter von Tschirnhaus[LL76])

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).

1.2 About the HTML versions of proofs

I have been experimenting with outputting formal proofs to HTML rather than LaTeX, in an effort to make reading them less effortful and tedious. The main benefits are these:

- Collapsible sections of text. This is helpful if you want to hide nasty parts of the proof by default, and for readers it is helpful for decluttering the screen once they are satisfied with the proof/justification of some lemma/claim, or once they have memorized the definition of a symbol.
- Pop-up references on cursor hover. This is more useful for the proofs that I do in this thesis than it usually is for proofs in mathematics, because of the much higher ratio of <# of fundamental symbols with no standard meaning> to <length of proof>.

The following features are planned, but may not be implemented until after my thesis is turned in.

[8] But see Section 3.5.1, for what I hope is only a temporary impediment

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

- Online multi-collaborator editing of proofs using the Google Realtime API.
- Maybe: option of using a completely-machine-readable language, hooked up to a first-order resolution theorem prover. I'm uncertain if anything besides satisfaction is gained by this.

2 Comparison to Related Work

2.1 Bayesian Reasoning

Bayesian reasoning and statistics feature prominently in most of my major examples. The Bayesian vs Frequentist debate is relevant here, and I recommend [Efr05] even for readers who are already familiar with it. What I plan to do with this section is explain how I formalize subjective probabilities, and compare my approach to others mentioned in the literature. I will introduce the problem of the interpretation of subjective probabilities, and explain why I cannot easily dismiss it, though I would like to. The purpose of this section is mainly to answer certain philosophical objections against the foundations of my work.

There are subjective assumptions that seem normal and obviously necessary, such as some of the assumptions I make in my arguments about assisted suicide, and more generally the kind assumptions one must make in order to derive anything with nontrivial ethical ramifications. And then there are subjective probabilities, which make most of us at least a little uneasy. I will argue that the origin of that uneasiness is the same as the origin of the uneasiness caused by adopting real-valued utility functions in utility theory, and then I will explain how I handle the two issues in the same way. The issue in both cases, I claim, is with the mixture of "qualitative" and "quantitative" subjective assumptions. My approach is to separate subjective probability assumptions into two parts: (1) a simple, subjective, qualitative part, and (2) a complex, objective, quantitative part. Doing that separation often requires the description of elaborate (but effectively-objective) hypotheticals, and experiments that will never be carried out, so I should reiterate that this is only a solution to the philosophical problem.

2.2 Truth in Mathematics / Intersubjective Agreement

This section will be about subjectiveness and vagueness in pure mathematics. I'll discuss some papers about the Continuum Hypothesis that I particularly like: Feferman's *Is the Continuum Hypothesis a Definite Mathematical Problem?* [Fef] and Koellner's response *Feferman on the Indefiniteness of CH.*

2.3 Coping with Vagueness

This will be a short section about some of the work on vagueness in philosophy, e.g. by Delia Fara. They like to talk about the Sorites Paradox. I'll show the vaguely-interpreted formal proofs way of handling it.

2.4 Logical Positivists

The logical positivists worked to varying degrees to expand the applications of formal logic to areas outside traditional mathematics. According to the most vocal of modern commentators, they were largely unsuccessful. It seems that the creation of a Leibnitzian dialog system (see Section 2.5) would have been considered very worthwhile among the logical positivists, and unlike Leibniz, they did have modern formal logic at their disposal. Given that, I attribute their lack of progress to two main factors:

1. Preoccupation with constructing elegant, widely-applicable theories. A premise of my approach is that an informal yes/no question must be fixed before formalization begins. Hence, the development of general theories about subjective and vague matters is explicitly not a goal; only making progress on the question matters. Of course, abstracting out common axiomatizations for reuse is still a good idea, but as with writing software libraries, it should not be done preemptively.
2. Working with examples outside of the problem domain I outlined in Section 1.1. Because there is no promise of discovering mathematically interesting material in the formal investigation of a question about a vague and subjective issue, the motivation for the very difficult work involved in rigorous reasoning must come entirely from elsewhere, namely from the question itself; we must be convinced that there is no easier way to make progress on the question, and that making progress on the question is indeed worth the work. Leibniz knew the danger of being insufficiently conscious of this point; speculating about why the project he envisioned had not been taken up by others earlier, like Descartes, he wrote:

   > *The true reason for this straying from the portal of knowledge is, I believe, that principles usually seem dry and not very attractive and are therefore dismissed with a mere taste.*
   >
   > (Gottfried Leibniz, 1679, *"On the General Characteristic"*[LL76])

   Also, the stigma in mathematics and science against working toward progress on value-laden issues has grown over time. Doing so gets one's work labeled as philosophy. The stigma is not surprising, given what has passed as good work in contemporary philosophy, but on the other hand, it is a clear fallacy of association to condemn a subject of study on account of the people who have managed, so far, to get payed to work on it.

2.5 Dialog Systems

A dialog system is a system of rules that two or more people/parties use to discuss or have some form of argument about a particular question. The following categories will help to convey the kind of dialog systems that I am interested in.

Terminology 1 (Practical dialog system)**.** In contrast to the work on *abstract argumentation frameworks*, which was popularized among computer scientists in [Dun95] and has sense spawned hundreds of publications[10] by nonmonotonic logic researchers, a **practical dialog system** is a dialog system that comes with non-trivial examples and guidelines for use, and which is not motivated by appeal to its generality or elegance. This is exemplified in the work of Douglas Walton and his colleagues (e.g [GW12], for a recent example) in the field of Informal Logic.

Terminology 2 (Leibnitzian dialog system)**.** A Leibnitzian dialog system is a practical dialog system whose purpose, above all else, is to provide a framework that suffices for rigorous, sound argumentation on an issue, to the fullest extent that such is possible, without being vulnerable to manipulation by fallacious reasoning or tactics of rhetoric. A Leibnitzian dialog system is fundamentally prescriptive, in contrast to the work of Walton and others, where the goal is to facilitate the construction of persuasive, defeasible reasoning arguments in a form that is (hopefully) more amenable to criticism.

I've chosen that tentative name due to writings by Leibniz, an idealist, who had a strong lifelong vision of such a system, but without the prerequisite of modern formal logic that is needed to concretely describe one. Writings like this:

> *For men can be debased by all other gifts; only right reason can be nothing but wholesome. But reason will be right beyond all doubt only when it is everywhere as clear and certain as only arithmetic has been until now. Then there will be an end to that burdensome raising of objections by which one person now usually plagues another and which turns so many away from the desire to reason. When one person argues, namely, his opponent, instead of examining his argument, answers generally, thus, 'How do you know that your reason is any truer than mine? What criterion of truth have you?' And if the first person persists in his argument, his hearers lack the patience to examine it. For usually many other problems have to be investigated first, and this would be the work of several weeks, following the laws of thought accepted until now. And so after much agitation, the emotions usually win out instead of reason, and we end the controversy by cutting the Gordian knot rather than untying it.*
>
> (Gottfried Leibniz, 1679, *"On the General Characteristic"*[LL76])

2.5.1 Abstract Argumentation Systems and Informal Logic Dialogue Systems

As I noted at the beginning of this section, there are two tracks of current work on dialog systems. The goal of the research in the first, which contains *abstract argumentation frameworks*, appears to be the development of descriptive mathematical models of (unrigorous) human argumentation, with the major motivation being to provide a theoretical framework for the academic analysis of natural language arguments. In the other, more-applied track, in which Douglas Walton does much of his work, the major motivation is the development of implementable computer tools to help people more-efficiently construct better and easier-to-criticize defeasible arguments.

[10] Google reports 1701 citations of that paper, as of 11 Aug 2012.

The goal of my work is closer to the second track, but more-prescriptive. I wish to develop a mathematical model for the informal concept of rigorous reasoning about subjective and vague questions (of the sort that I described in Section 1.1), where I mean "rigorous" to imply exceptional internal consistency, transparency, and clarity – properties viewed as virtuous among sober and unattached intellectuals.

2.6 Defeasible, Nonclassical, and Intensional Logics

The main purpose of this section is to explain why I stick to a version of classical FOL.

Let L be a nonclassical logic or intensional logic. I claim that the ease of converting proofs in L to equally-readable proofs in a user-friendly version of FOL (Section 3.1), or of extending the definition of user-friendly FOL to accommodate the useful features of L (which I've already done several times), is inversely proportional to the difficulty (relative to FOL) of interpreting L-sentences. Furthermore, when there is no added difficulty of interpreting sentences, or when the features of L allow us to write *easier* to interpret sentences (and without making it too much easier to write deceptively-simple sentences whose meaning is complex), then we can either already simulate those features with low overhead in user-friendly FOL, or else we can extend our definition of user-friendly FOL to accommodate those features.

This will not be an argument against the study of defeasible, nonclassical, and intentional logics in general, because its force depends on two uncommon aspects of my project: (1) sentences about vague, subjective concepts and uncertain knowledge are already more-difficult to interpret than sentences about traditional mathematical concepts and certain knowledge, and (2) with the minimally-reductionist approach to proofs that I advocate, ease of interpretation is more important than it usually is in applications of formal logic.

Also see Section 3.4.

3 Framework

3.1 Logic Definitions

This section gives a version of sorted first-order logic with subtyping, partial functions/undefinedeness, variable arity overloading, and sort operators, which I'll refer to as $\mathsf{MSFOL}_{\subseteq,\perp}$ (many-sorted FOL with subtyping and undefinedness). The partial functions feature is based on [Far93].

Definition 1 ($\mathcal{S}$-type, $\mathcal{S}$-sort, $\mathcal{S}$-predicate type, etc)**.** When $\mathcal{S}$ is a set of symbols, each of which is designated a sort symbol or sort-operator symbol, and such that $\mathcal{S}$ is equipped with a function that assigns a fixed arity $\geqslant 1$ to each sort operator symbol, then the $\mathcal{S}$-types are defined by the following grammar:[11]

sort	S	$:=$	$\langle$sort-symbol in $\mathcal{S}\rangle \mid \langle$sort-operator symbol in $\mathcal{S}\rangle(S^+) \mid S \cap S \mid S \backslash S \mid S \cup S$
domain type	D	$:=$	S^+
simple total function type		$:=$	$D \to S$
simple predicate type		$:=$	$D \to \mathbb{B}$
simple partial function type		$:=$	$D \to_? S$
function type		$:=$	nonempty set of simple total function types
partial function type		$:=$	nonempty set of simple total and partial function types
predicate type		$:=$	nonempty set of simple predicate types, or $\mathbb{B}$

Definition 2 (arity)**.** The *arity* of a domain type $S_1 \times \cdots \times S_k$ is k. The arity of a simple (total or partial) function type or simple predicate type is the arity of its domain type.

Definition 3 (signature)**.** A $\mathsf{MSFOL}_{\subseteq,\perp}$ signature Σ is given by:

- A set $\mathcal{S}$ of *sort symbols* and *sort operator symbols* . A nonempty subset of the sort symbols are designated *top-level sort symbols*, and each sort operator symbol has a fixed arity $\geqslant 1$.
- A set of *term symbols*, each of which is assigned an $\mathcal{S}$-type. The term symbols are partitioned into three sets:
  - Constant symbols, which are assigned sorts.
  - Function symbols, which are assigned partial function types.
  - Predicate symbols, which are assigned predicate types.
- A set of sort constraints each of the form $S_1 \subseteq S_2$ where S_1, S_2 are $\mathcal{S}$-sorts.

Definition 4 (Σ-structure)**.** A Σ-structure $\mathcal{M}$ for a signature Σ consists of a set $\underline{\mathcal{M}}$ called the universe of $\mathcal{M}$ and a mapping from various pieces of syntax Λ to elements of $\underline{\mathcal{M}} + \perp$, subsets of $\underline{\mathcal{M}}$, partial functions from $\underline{\mathcal{M}}^*$ to $\underline{\mathcal{M}}$, relations on $\underline{\mathcal{M}}^*$, or truth values, that meets the following constraints, where regardless of the kind of syntax that Λ is, we write $\Lambda^{\mathcal{M}}$ for the object that $\mathcal{M}$ assigns to Λ. Partial functions on $\underline{\mathcal{M}}$ are total functions on $\underline{\mathcal{M}} + \perp$, where $\perp$ is a new object not in $\underline{\mathcal{M}}$.

- For each sort symbol s, $s^{\mathcal{M}}$ is a nonempty subset of the universe $\underline{\mathcal{M}}$, and if $s_1, \ldots, s_k$ are the top-level sort symbols then $s_1^{\mathcal{M}}, \ldots, s_k^{\mathcal{M}}$ is a partition of the universe.
- For each sort operator symbol F with arity n, $F^{\mathcal{M}}$ is a function from $(2^{\mathcal{M}})^n$ to $2^{\mathcal{M}}$.
- If $S = S_1 \backslash S_2$ then $S^{\mathcal{M}}$ is the set difference of $S_1^{\mathcal{M}}$ and $S_2^{\mathcal{M}}$, i.e. the elements of S_1 that are not in S_2. Likewise $\cap$ and $\cup$ have their expected meanings.

[11] I write X^+ to mean a nonempty list of Xs, and alternation | has the same meaning as in regular expressions. When I write $\langle$sort-operator symbol in $\mathcal{S}\rangle(S^+)$, I of course mean that the number of sort arguments should match the arity of the sort-operator symbol.

- If $S = F(S_1, \ldots, S_n)$ for F a sort-operator symbol of arity n then $S^{\mathcal{M}} = F^{\mathcal{M}}(S_1^{\mathcal{M}}, \ldots, S_n^{\mathcal{M}})$.
- If $S_1 \sqsubseteq S_2$ is a sort constraint then $S_1^{\mathcal{M}}$ is a subset of $S_2^{\mathcal{M}}$.
- If $D = X_1, \ldots, X_k$ is a domain type, then $D^{\mathcal{M}}$ is the set $X_1^{\mathcal{M}} \times \cdots \times X_k^{\mathcal{M}}$. Also $\mathbf{D}^{\mathcal{M}+\perp}$ is the set $(X_1^{\mathcal{M}} + \perp) \times \cdots \times (X_k^{\mathcal{M}} + \perp)$.
- If c is a constant with type S then $c^{\mathcal{M}} \in S^{\mathcal{M}}$.
- If f is a function symbol with the following partial function type

  $$\begin{array}{ll} \{ & D_1 \to S_1, \ldots, D_k \to S_k, \\ & D_{k+1} \to_? S_{k+1}, \ldots, D_n \to_? S_n \quad \} \end{array}$$

  then $f^{\mathcal{M}}$ is a function with domain $D_1^{\mathcal{M}+\perp} \cup \ldots \cup D_n^{\mathcal{M}+\perp}$ such that
  - for each $i \in \{1, \ldots, k\}$ if $\vec{a} \in D_i^{\mathcal{M}}$ (so no component is $\perp$) then $f^{\mathcal{M}}(\vec{a}) \in S_i^{\mathcal{M}}$.
  - for each $i \in \{k+1, \ldots, n\}$ if $\vec{a} \in D_i^{\mathcal{M}}$ (so no component is $\perp$) then $f^{\mathcal{M}}(\vec{a}) \in S_i^{\mathcal{M}} + \perp$.
  - if $\vec{a}$ is in the domain of $f^{\mathcal{M}}$ and some component of $\vec{a}$ is $\perp$, then $f^{\mathcal{M}}(\vec{a}) = \perp$.
- If P is a predicate symbol with type $D_1 \to \mathbb{B}, \ldots, D_n \to \mathbb{B}$ then $P^{\mathcal{M}}$ is a relation with domain $D_1^{\mathcal{M}+\perp} \cup \ldots \cup D_n^{\mathcal{M}+\perp}$ such that if $\vec{a}$ is in the domain of $P^{\mathcal{M}}$ and some component of $\vec{a}$ is $\perp$, then $P^{\mathcal{M}}(\vec{a})$ is false.[12] When $n = 0$, $P^{\mathcal{M}} \in \{\text{true}, \text{false}\}$.

Grammar for terms and formulas for a given signature Σ.

variable	x		
sort	S		
constant symbol	c		
(partial) function symbol	f		
predicate symbol	P		
proper term	t	$\coloneqq$	$c \mid x \mid f(t^+)$
atomic formula	R	$\coloneqq$	$P \mid P(t^+) \mid t = t \mid t{\downarrow} \mid S(t)$
formula	A	$\coloneqq$	$R \mid A \wedge A \mid A \vee A \mid \neg A \mid \forall x{:}S.A \mid \exists x{:}S.A$
term	Λ	$\coloneqq$	$t \mid A$

Evaluation of terms and formulas

Definition 5 (context, object assignment)**.** Let Σ be a signature and $\mathcal{M}$ a Σ-structure. An $\mathcal{M}$-object assignment is, as usual, a mapping from some finite set of variables to elements of the universe $\underline{\mathcal{M}}$. A Σ-context (or just context, when unambiguous) is a finite mapping from variables to sorts. For κ a context, a $\langle \kappa, \mathcal{M} \rangle$-object assignment σ is an object assignment whose domain is the domain of κ such that $\sigma(x)$ is an element of $(\kappa(x))^{\mathcal{M}}$ for all x.

We are now ready to define the evaluation of terms over a given structure. Fix a signature Σ and Σ-structure $\mathcal{M}$. The valuation function for $\mathcal{M}$ is a partial function whose domain is

[12] Note that a relation is given by two sets: the relation's domain, and the set of objects that hold for the relation, which is a subset of the domain.

the set of pairs (term, object assignment) and whose range is $\underline{\mathcal{M}} + \bot$. We use the normal notation $t^{\mathcal{M}}[\sigma]$, $A^{\mathcal{M}}[\sigma]$. **There are two notions of "defined" involved here.** When $\mathcal{M}$'s valuation function is undefined on a (term, object assignment) pair $\langle \Lambda, \sigma \rangle$, we will say that the evaluation $\Lambda^{\mathcal{M}}[\sigma]$ crashes or that $\mathcal{M}, \sigma$ crashes Λ. In that case Λ is not a meaningful term with respect to $\mathcal{M}$ and σ, and we will use this notion to define what counts as a syntax error. The other notion is when $\Lambda^{\mathcal{M}}[\sigma] = \bot$, in which case Λ *is* a meaningful term with respect to $\mathcal{M}$ and σ; this notion is analogous to the "exception" feature found in many modern programming languages.

In the following recursive definition, we keep implicit the following constraint: $\mathcal{M}$'s valuation function on $\langle \Lambda, \sigma \rangle$ is undefined (crashes) if evaluation crashes at any subterm. When $\vec{t} = t_1, \ldots, t_k$ we make $\vec{t}^{\mathcal{M}}[\sigma]$ denote $t_1^{\mathcal{M}}[\sigma], \ldots, t_k^{\mathcal{M}}[\sigma]$.

Proper terms:

- $x^{\mathcal{M}}[\sigma] = \sigma(x)$ and evaluation crashes if x is not in the domain of σ.
- $c^{\mathcal{M}}[\sigma] = c^{\mathcal{M}}$
- $f(\vec{t})^{\mathcal{M}}[\sigma] = f^{\mathcal{M}}(\vec{t}^{\mathcal{M}}[\sigma])$. Evaluation crashes if $\vec{t}^{\mathcal{M}}[\sigma]$ is in $\underline{\mathcal{M}}^*$ but not in the domain of $f^{\mathcal{M}}$.

Formulas:

- $(t_1 = t_2)^{\mathcal{M}}[\sigma]$ holds iff $t_1^{\mathcal{M}}[\sigma]$ and $t_2^{\mathcal{M}}[\sigma]$ are the same element of the universe $\underline{\mathcal{M}}$.[13]
- $P^{\mathcal{M}}[\sigma] = P^{\mathcal{M}}$ and evaluation crashes if the type of P is not $\mathbb{B}$.
- $P(\vec{t})^{\mathcal{M}}[\sigma]$ holds iff $\vec{t}^{\mathcal{M}}[\sigma]$ is in $P^{\mathcal{M}}$. Evaluation crashes if $\vec{t}^{\mathcal{M}}[\sigma]$ is in $\underline{\mathcal{M}}^*$ but not in the domain of $P^{\mathcal{M}}$.[14]
- $(A \wedge B)^{\mathcal{M}}[\sigma]$ holds iff $A^{\mathcal{M}}[\sigma]$ holds and $B^{\mathcal{M}}[\sigma]$ holds. Similarly for $\vee, \neg, \Rightarrow$.
- To evaluate $(\forall x{:}S.A)^{\mathcal{M}}[\sigma]$ or $(\exists x{:}S.A)^{\mathcal{M}}[\sigma]$, first evaluate $A^{\mathcal{M}}[\sigma, x \mapsto a]$ for every $a \in S^{\mathcal{M}}$. If $S^{\mathcal{M}}$ is empty, then evaluation crashes at this level. If no subevaluation has crashed yet, then evaluation does not crash at this level, and in that case the definition is as usual:
  - $(\forall x{:}S.A)^{\mathcal{M}}[\sigma]$ evaluates to true iff $A^{\mathcal{M}}[\sigma, x \mapsto a]$ evaluates to true for every $a \in S^{\mathcal{M}}$.
  - $(\exists x{:}S.A)^{\mathcal{M}}[\sigma]$ evaluates to true iff $A^{\mathcal{M}}[\sigma, x \mapsto a]$ evaluates to true for some $a \in S^{\mathcal{M}}$.
- $t{\downarrow}^{\mathcal{M}}[\sigma]$ holds iff $t^{\mathcal{M}}[\sigma]$ is in $\underline{\mathcal{M}}$, i.e. if $t^{\mathcal{M}}[\sigma] \neq \bot$.
- $(S(t))^{\mathcal{M}}[\sigma]$ holds iff $t^{\mathcal{M}}[\sigma]$ is in $S^{\mathcal{M}}$.

Definition 6 (consistent signature)**.** A signature Σ is *consistent* iff the set of Σ-structures is non-empty.

It is not hard to define an inconsistent signature using just sort constraints. It is also possible for a signature to be inconsistent without the sort constraints being inconsistent, or with no sort constraints:

[13] Hence $=$ is untyped. Also note that $\bot =^{\mathcal{M}} \bot$ is false.

[14] Recall that $\vec{a} \in P^{\mathcal{M}}$ implies $\vec{a} \in \underline{\mathcal{M}}^*$, and hence if any of the $t_i^{\mathcal{M}}[\sigma]$ are error values, and none causes a crash, then $P(\vec{t})^{\mathcal{M}}[\sigma]$ is defined and false.

Example The signature with top-level sorts S_1, S_2 and a single function symbol of type $\{S_1 \to S_1 \cap S_2\}$ is inconsistent.

Definition 7 (well-typed terms)**.** For Σ a consistent signature, a Σ-term Λ is ill-typed in context κ iff there is a Σ-structure $\mathcal{M}$ and $\langle \kappa, \mathcal{M} \rangle$-object assignment σ that makes the evaluation $\Lambda^{\mathcal{M}}[\sigma]$ crash. Λ is well-typed in κ iff it is not ill-typed in κ.

The problem of determining whether a given signature is consistent is decidable (by Theorems 1-2). The problem of determining, given $\langle \Sigma, A \rangle$, whether A is a well-typed Σ-sentence, is also decidable (by Theorems 1-3). For a given signature Σ, the set of valid Σ-sentences is recursively enumerable (by Theorems 1 and 4). Proofs of these claims will appear in my dissertation. They are tedious, but not difficult.

Definition 8 ($\mathsf{MSFOL}_{\subseteq}$)**.** $\mathsf{MSFOL}_{\subseteq}$ is the simplified version of $\mathsf{MSFOL}_{\subseteq,\perp}$ that has no partial function types or undefinedness. Every function is total on its domain, and there is no $\downarrow$ symbol.

Theorem 1. *There is a polynomial time reduction that, given a $\mathsf{MSFOL}_{\subseteq,\perp}$ signature Σ and a Σ-sentence A, produces a $\mathsf{MSFOL}_{\subseteq}$ signature Σ' and Σ'-sentence A' such that:*

- *Σ is consistent iff Σ' is.*
- *A is a well-typed Σ-sentence iff A' is a well-typed Σ'-sentence.*
- *A is a valid Σ-sentence iff A' is a valid Σ'-sentence.*

Theorem 2. *The problem of determining whether a given $\mathsf{MSFOL}_{\subseteq}$ signature is consistent is reducible to the validity problem for monadic predicate logic*[15]*, and thus is decidable..*

Theorem 3. *Given a $\mathsf{MSFOL}_{\subseteq}$ signature Σ and Σ-sentence A, the problem of determining if A is a well-typed Σ-sentence is reducible to the validity problem for monadic predicate logic, and thus is decidable.*

Theorem 4. *For Σ a consistent $\mathsf{MSFOL}_{\subseteq}$ signature, the validity problem for well-typed Σ-sentences is reducible to the validity problem for FOL.*

3.1.1 Reductions and proof sketches for appendix

Definition 9. For a $\mathsf{MSFOL}_{\subseteq}$ signature Σ, if ρ is a function (resp. predicate) symbol of Σ and k is a positive integer then $\mathsf{type}_{\Sigma,k}(\rho)$ is the set of arity-k simple function (resp. predicate) types of ρ. The argument Σ is left out when it's clear from the context.

Reduction for Theorem 1

The reduction is given by a polynomial time function that, given a $\mathsf{MSFOL}_{\subseteq,\perp}$ signature Σ and a (not necessarily well-typd) Σ-formula A, yields a $\mathsf{MSFOL}_{\subseteq}$ signature Σ', a set X of Σ'-sentences, and a Σ'-formula A'.

[15] i.e. predicate logic with only unary predicate symbols and constants

We modify Σ to get Σ', and generate the set X at the same time. Add a new top-level sort symbol Undef, and a new constant $\bot : \mathsf{Undef}$. Add $\forall x{:}\mathsf{Undef}.\, x = \bot$ to X. For each sort symbol s of Σ, add the constraint $s \cap \mathsf{Undef} = \varnothing$. Now we modify each of the simple function or predicate types of each (partial) function or predicate symbol ρ. Recall that the type of a (partial) function symbol is a set of simple total function types (resp. simple total and partial function types), and similarly for predicate symbols. For notational convenience, I will just show what to do for types of arity 2; it will be obvious what to do for smaller and larger arities. When S is a sort, let $S^\bot$ abbreviate $S \cup \mathsf{Undef}$.

- If $S_1 \times S_2 \to S$ is a simple function type of ρ, then add to it the two simple function types $\{\mathsf{Undef} \times S_2^\bot \to \mathsf{Undef}, S_1^\bot \times \mathsf{Undef} \to \mathsf{Undef}\}$, which force the function to be undefined whenever any of its arguments are undefined.
- If $S_1 \times S_2 \to_? S$ is a simple partial function type of ρ, then replace it with the three simple function types $\{\mathsf{Undef} \times S_2^\bot \to \mathsf{Undef}, S_1^\bot \times \mathsf{Undef} \to \mathsf{Undef}, S_1^\bot \times S_2^\bot \to S^\bot\}$. This accomplishes the same thing as we did for total function types, except that it allows the partial function to be undefined when all of its arguments are defined.
- If $S_1 \times S_2 \to \mathbb{B}$ is a simple predicate type of ρ, then replace it with $S_1^\bot \times S_2^\bot \to \mathbb{B}$. Also add to X the sentence $\forall x_1{:}S_1^\bot.\, \forall x_2{:}S_2^\bot.\, (x_1 = \bot \vee x_2 = \bot) \Rightarrow \neg\rho(x_1, x_2)$, which forces the relation to be false whenever any of its arguments are undefined.

Next we modify the Σ-sentence A to get the Σ'-sentence A'. Let A'' be obtained by replacing each subformula of the form $t_1 = t_2$ with $(t_1 \neq \bot \wedge t_2 \neq \bot \wedge t_1 = t_2)$, and each subformula of the form $t{\downarrow}$ with $t \neq \bot$. Then A' is $(\bigwedge_{B \in X} B) \to A''$.

To prove correctness of the reduction (i.e. Theorem 1), define a bijection between Σ-structures, and Σ'-structures that satisfy X. The sentences X are used to translate Σ'-structures to Σ-structures, since Σ' alone does not force structures to interpret Undef as a 1-element set, or to make relations false whenever any of the arguments is the interpretation of $\bot$.

Reduction for Theorems 2 and 3

Fix a $\mathsf{MSFOL}_\subseteq$ signature Σ. Let $\mathcal{L}$ be the monadic predicate logic language consisting of the sort symbols of Σ as unary predicate symbols. If S is a sort of Σ then $\ulcorner S \urcorner$ is a function that maps strings to strings, such that $\ulcorner S \urcorner(x)$ is an $\mathcal{L}$-formula containing the variable x, and $\ulcorner S \urcorner(t)$ is $\ulcorner S \urcorner(x)[x \mapsto \ulcorner t \urcorner]$.

TODO: sort operators

$$\begin{array}{rcll} \ulcorner s \urcorner &=& x \mapsto s(x) & \text{for } s \in \mathcal{L} \\ \ulcorner S_1 \cap S_2 \urcorner &=& x \mapsto \ulcorner S_1 \urcorner(x) \wedge \ulcorner S_2 \urcorner(x) & \\ \ulcorner S_1 \cup S_2 \urcorner &=& x \mapsto \ulcorner S_1 \urcorner(x) \vee \ulcorner S_2 \urcorner(x) & \\ \ulcorner S_1 \backslash S_2 \urcorner &=& x \mapsto \ulcorner S_1 \urcorner(x) \wedge \neg \ulcorner S_2 \urcorner(x) & \\ \ulcorner \varnothing \urcorner &=& x \mapsto x \neq x & \end{array}$$

Additionally,

- $\ulcorner \cdot \urcorner$ maps each sort constraint $S_1 \subseteq S_2$ of Σ to the $\mathcal{L}$-sentence $\forall x.\, \ulcorner S_1 \urcorner(x) \Rightarrow \ulcorner S_2 \urcorner(x)$.

- If $D = S_1 \times \cdots S_k$ is an arity k domain type then $\ulcorner D \urcorner$ is the function that takes any k strings $u_1, \ldots, u_k$ to $\ulcorner S_1 \urcorner(u_1) \wedge \ldots \wedge \ulcorner S_k \urcorner(u_k)$ (which is an $\mathcal{L}$-formula if the u_i are variables).
- If $D_1, \ldots, D_m$ are the arity-k domain types of function or predicate symbol ρ, then let $\mathsf{dom}_{\rho,k}$ be the function that takes any k strings $\vec{u}$ to $\ulcorner D_1 \urcorner(\vec{u}) \vee \ldots \vee \ulcorner D_m \urcorner(\vec{u})$.
- If $D_1 \to S_1, \ldots, D_m \to S_m$ are the arity-k simple function types of function symbol f, then let $\mathsf{range}_{f,k}$ be the function that takes any $k+1$ strings $u_1, \ldots, u_k, u_{k+1}$ to

$$\bigwedge_{i \leqslant m} \ulcorner D_i \urcorner(x_1, \ldots, x_k) \Rightarrow \ulcorner S_i \urcorner(x_{k+1})$$

which is an $\mathcal{L}$-formula if the u_i are variables.

The output $\psi = \psi_\sigma$, a monadic predicate logic $\mathcal{L}$-sentence, of the reduction for Theorem 2, is give by the conjunction of the $\mathcal{L}$-sentences:

- $\ulcorner S_1 \subseteq S_2 \urcorner$ for each sort constraint.
- $\exists x.\, s(x)$ for each sort symbol.
- $\exists x.\, \ulcorner S \urcorner(x)$ for each constant c of type S.
- For each function symbol f, and each k such that f has a simple function type of arity k, if $\{D_1 \to S_1, \ldots, D_m \to S_m\}$ are those simple function types, then for every nonempty $Z \subseteq \{1, \ldots, m\}$, the sentence

$$\left(\exists \vec{x}.\, \bigwedge_{i \in Z} \ulcorner D_i \urcorner(\vec{x})\right) \Rightarrow \left(\exists x.\, \bigwedge_{i \in Z} \ulcorner S_i \urcorner(x)\right)$$

It then suffices to prove that Σ is consistent iff ψ is satisfiable.

The $\Rightarrow$ direction is easy - just take the most obvious approach of having the universe of the Σ structure $\mathcal{M}$ be the universe of the ψ model $\mathcal{N}$, and having $\mathcal{N}$'s relations be the same as $\mathcal{M}$'s interpretations of its sort symbols. Then verify that $\mathcal{N}$ satisfies ψ.

For the $\Leftarrow$ direction, let $\mathcal{N}$ be a model of ψ. We construct a Σ-structure $\mathcal{M}$. The universe U of $\mathcal{M}$ is the closure of $\{s^{\mathcal{N}} \mid s \in \text{ sort symbol of } \Sigma\}$ under intersection, union, and set difference, *minus* the empty set. Each predicate symbol of Σ gets interpreted as the relation that is true for every element of its domain. If the sort of constant c of Σ is S, then $c^{\mathcal{M}} = \{a \in U \mid (\ulcorner S \urcorner(x))^{\mathcal{N}}[x \mapsto a]\}$. Now let f be a function symbol of Σ. We have already determined the domain of f. For $\vec{a}$ in the domain of f, with $k = |\vec{a}|$, let I be the indices of f's arity-k simple function types. Let $I' \subseteq I$ be the $i \in I$ such that $D_i \to S_i$ is the i-th such simple function type of arity k and $\vec{a} \in D_i^{\mathcal{M}}$. Then $f(\vec{a})$ is $V = \bigcap_{i \in I'} \{a \in U \mid (\ulcorner S_i \urcorner(x))^{\mathcal{N}}[x \mapsto a]\}$. Note that V must be in the universe of $\mathcal{M}$ that we specified, since it is obtainable from the atomic relations $s^{\mathcal{N}}$ using intersection, union and set difference, and because it is not the empty set by virtue of $\mathcal{N}$ satisfying the $\mathcal{L}$-axioms for function symbols.

Now we give the reduction for Theorem 3, which uses the reduction for Theorem 2, Let A be a Σ-sentence. Rename bound variables so that no variable is bound by more than one quantifier. Introduce a variable y_t for each subterm t of A. Now we extend the definition of $\ulcorner\cdot\urcorner$ to (the subterms of) A:

$$\begin{array}{rcll}
\ulcorner A_1 \text{ op } A_2 \urcorner & = & \ulcorner A_1 \urcorner \wedge \ulcorner A_2 \urcorner & \text{for op} \in \{\wedge, \vee, \Rightarrow\} \\
\ulcorner \neg A_1 \urcorner & = & \ulcorner A_1 \urcorner & \\
\ulcorner Qx{:}S.\, A \urcorner & = & \ulcorner S \urcorner(y_x) \Rightarrow \ulcorner A \urcorner & \text{for } Q \in \{\forall, \exists\} \\
\ulcorner x \urcorner & = & y_x & \\
\ulcorner c \urcorner & = & y_c & \\
\ulcorner f(t_1, \ldots, t_k) \urcorner & = & \mathsf{dom}_{f,k}(y_{t_1}, \ldots, y_{t_k}) \wedge \mathsf{range}_{f,k}(y_{t_1}, \ldots, y_{t_k}, y_{f(t_1,\ldots,t_k)}) & \\
\ulcorner P(t_1, \ldots, t_k) \urcorner & = & \mathsf{dom}_{P,k}(y_{t_1}, \ldots, y_{t_k}) & \\
\ulcorner t_1 = t_2 \urcorner & = & \ulcorner t_1 \urcorner \wedge \ulcorner t_2 \urcorner &
\end{array}$$

Let $c_1, \ldots, c_n$ be the constants of Σ of sorts $S_1, \ldots, S_n$ that are used in A. Then the final monadic predicate logic $\mathcal{L}$-sentence is obtained by universally quantifying over all the free variables in

$$\psi \wedge \ulcorner S_1 \urcorner(y_{c_1}) \wedge \ldots \wedge \ulcorner S_n \urcorner(y_{c_n}) \Rightarrow \ulcorner A \urcorner$$

Reduction for Theorem 4 Let Σ be a consistent $\mathsf{MSFOL}_{\subseteq}$ signature. Let $\mathcal{L} = \mathcal{L}_\Sigma$ be the first-order language with:

- the same constant symbols as Σ plus one additional constant none.
- a unary predicate symbol for every sort symbol of Σ.
- for each function (resp. predicate) symbol ρ, a k-ary function (resp. predicate) symbol ρ_k for every k such that ρ's type has at least one simple function (resp. predicate) type of arity k (equivalently, $\mathsf{type}_k(\rho)$ is non-empty).

Let Γ be the following set of $\mathcal{L}$-sentences. The conjunction of its elements will be in the antecedent of the reduction's final $\mathcal{L}$-sentence. Below, $\vec{x}$ abbreviates $x_1, \ldots, x_k$.

- For each of $\mathcal{L}$'s function symbols f_k the sentences

$$\forall \vec{x}.\ \mathsf{range}_{f,k}(\vec{x}, f_k(\vec{x}))$$

$$\forall \vec{x}.\ f_k(\vec{x}) \neq \mathsf{none} \Leftrightarrow \mathsf{dom}_{f,k}(\vec{x})$$

- For each of $\mathcal{L}$'s predicate symbols P_k the sentence

$$\forall \vec{x}.\ P_k(\vec{x}) \Rightarrow \mathsf{dom}_{P,k}(\vec{x})$$

- For each constant c in Σ of sort S, the sentence $\ulcorner S \urcorner(x)$.
- For each sort symbol s, the sentence $\exists x.s(x)$.
- For each sort constraint $S_1 \subseteq S_2$, the sentence $\ulcorner S_1 \subseteq S_2 \urcorner$.
- For $s_1, \ldots, s_k$ the top-level sort symbols of Σ, the following two sentences, which express that for any model $\mathcal{N}$, the $k+1$ sets $s_1^{\mathcal{N}}, \ldots, s_k^{\mathcal{N}}, \{\mathsf{none}^{\mathcal{N}}\}$ are a partition of the universe.

$$\forall x.\ (\neg s_1(x) \wedge \ldots \wedge \neg s_k(x)) \Leftrightarrow x = \mathsf{none}$$

$$\forall x.\ \bigwedge_{i<j\leqslant k} \neg s_i(x) \vee \neg s_j(x)$$

Let A be a well-typed Σ-sentence. Obtain an $\mathcal{L}$-sentence ϕ by modifying A according to the following rules until no more such modifications are possible:

- Subformula $\forall x{:}S.\, A'$ becomes $\forall x.\ulcorner S \urcorner(x) \Rightarrow A'$.
- Subformula $\exists x{:}S.\, A'$ becomes $\exists x.\ulcorner S \urcorner(x) \wedge A'$.
- Subterm $f(t_1, \ldots, t_k)$ becomes $f_k(t_1, \ldots, t_k)$.
- Subformula $P(t_1, \ldots, t_k)$ becomes $P_k(t_1, \ldots, t_k)$.

The final $\mathcal{L}$-sentence returned by the reduction is $\left(\bigwedge_{\psi \in \Gamma} \psi \right) \Rightarrow \phi$. If π is that sentence, then it satisfies: A is a valid Σ-sentence iff π is a valid $\mathcal{L}$-sentence.

3.2 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 for an author of a proof to describe, in the metalanguage, their intended semantics. In my thesis I will use Douglas Walton's fallacious proof about marriage ("nobody should ever get married") to illustrate the definitions in this section.

This definition of *vaguely-interpreted formal proof* in this section is tailored for the particular variant of sorted first-order logic that I use in the examples and define in Section 3.1, 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 classical first order logic.

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 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 quite vague. The author 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. The author should try to minimize the use of simplifying assumptions.
- 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 *a* "standard interpretation" of the symbol; due to the vagueness in the problems we are interested in, a set of axioms will have many standard models. Typically $g(s)$ will be between the length of a sentence and a long paragraph, but can be longer.

[16]Recall that a signature's language has *sort symbols*, which structures must interpret as subsets of the universe, and a signature may constrain their interpretations by *sort constraints*, which are expressions of forms like $S_1 \subseteq S_2$ (meaning that the interpretation of S_1 must be a subset of the interpretation of S_2) or $S_1 \cap S_2 = \varnothing$, etc. A language can also have *sort operator symbols*, which are second order function symbols that can only be applied to sorts. In this thesis sort operators have a 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).[17] A signature assigns sorts to its constants, and types to its function and predicate symbols. In this thesis, 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 usually-uninteresting ones (namely, typing axioms and axioms that express relationships between sorts).

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 language is just a set of symbols, each of which is designated a constant, predicate, function, sort, or sort-operator symbol.

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 and sort constraint 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 $\mathcal{L}_{\mathsf{def}}$. Func-

[16] I may end up moving this paragraph.

[17] 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.

tion 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,[18] and $\mathcal{L}_{\mathsf{def}}$ contains neither sort nor sort-operator symbols[19].
- 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 $\Gamma_{\mathsf{assum}} \cup \Gamma_{\mathsf{simp}} \cup \Gamma_{\mathsf{claim}} \cup \Gamma_{\mathsf{defn}}$.
- 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

[18] 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.

[19] 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

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 ideal 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.

3.2.1 My prescription for formal mathematics (with 5 color theorem example and failed proof of 4 color theorem)

I have an unusual opinion about formal mathematics: I agree with the community that it is worthwhile to try to make formal theorem proving easy enough that it can one day improve the trustworthiness of proofs in mathematics (especially in computer science), but on the other hand I think that the way it is being done right now often makes poor use of time.

I'll call my prescription *minimally-reductionist* formal theorem proving (or minimally-reductionist formal mathematics). It uses a notion of proofs that is based on the definition of vaguely-interpreted formal proofs, with the following roughly-defined additions:

1. There is a language of structured machine-checkable proofs, of the same style as in Mizar.

2. We add the sentence label lem for lemmas and theorems that are accompanied by machine-checkable proofs.

3. Declarations of new symbols and sorts are allowed throughout the proof.

There is no syntactic guarantee that the axioms used in a proof are consistent, as in existing interactive theorem proving systems. Instead, the language interpretation guide is relied upon for verifying consistency (informally), and more-generally for verifying that the axioms are sound with respect to their intended semantics (which existing interactive theorem proving systems must do informally as well).

The greatest difference of my prescription is in the recommended use of the system. Not only is it not necessary to define objects in terms of simpler or more-fundamental objects, it is actually discouraged unless the definition is used in the proof. Furthermore, authors are encouraged to introduce fundamental symbols for things that are not literally assigned to symbols in the prose source proof. As an example, in my formal proof of the 5 color theorem, there is a 6-argument symbol for "nodes u_1, u_2, u_3, u_4, u_5 are arranged clockwise around node v". It is easy to say what it means in prose, and we only need to use one proposition about it, so it is a waste of time to define it formally.

The current draft of the formal proof of the 5-color theorem can be found here.

3.3 Theoretical Dialogue System

3.3.1 Vaguely-Interpreted Formal Proofs with Finite Models

For now I will just give a preview of the definition of a state of a dialog. It includes:

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.

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 will sometimes require introducing new versions of sort and term symbols in $\mathcal{L}_{\mathsf{math}}$ that have finite intended interpretations. For example, in the Berkeley argument from Section 4, 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 the dialog system *theoretical*.

In my thesis, 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 1.1), doing so is always possible *in principle*.

3.3.2 Dialogue rules and definition of progress

The rules that I'll define precisely here will consist of:

1. Rules for moves similar to those given at the end of Section 3.2 for criticizing a vaguely-interpreted formal proof.
2. Rules for moves that extend the language.

It turns out to be easy to ensure progress if the language is not extended indefinitely. I 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}.[20] We should also allow each party to make a bounded number of language-extending moves that do not make progress.

3.3.3 Example using Sue Rodriguez argument

I will sketch what a dialogue could look like where one party is carrying out the argument introduced in Section 7.1 and the other is criticizing.

3.4 Patterns for formalizing common kinds of defeasible reasoning

When I explain my work to intelligent people outside of mathematical disciplines –students of law, politics, philosophy; incidentally the people who take the greatest interest in it– the most difficult task is explaining the practical effect of the limitation to deductive reasoning.

Consider the following excerpt from a recent paper [Wal11] by Walton, which advocates the use of defeasible logic:

> *The most widely useful argumentation schemes that fit arguments in everyday conversational argumentation are defeasible ones* [citation omitted]*. A good example is argument from expert opinion. This scheme is not well modeled by a deductive interpretation. Basing it on an absolutely universal generalization, to the effect that what an expert says is always true, does not yield a useful logical model. Indeed such a deductive model would make the scheme into a fallacious form of argument by making it unalterably rigid. In practice, evaluating an argument from expert opinion is best carried out by seeing how well it survives the testing procedure of critical questioning* [citation omitted].

Ironically, Walton has employed a straw man argument here (but on the other hand, he does say that he is talking about "everyday conversational argumentation", which is certainly not my interest). This is typical in the motivation for defeasible logic: implicitly the proponent suggests that in a deductive logic framework, to formalize an argument from expert opinion,

[20] 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

or an instance of inductive or abductive reasoning, or an instance of default reasoning, the only option is to assert a general rule that has obvious fallacious instances. A slightly more dignified criticism of deductive logic makes the implicit suggestion that in a deductive framework, the only option is to formulate a concise and elegant schema for some type of defeasible argument, which suffices to justify all and only the good instances of that argument type. I say that it is slightly more dignified because it is sometimes an innocent instance of academics' often-very-productive instinct to generalize, to obsess about elegance and wide applicability.

In contrast, for arguing about contentious and important issues, when plenty of time is available, I think it is a good idea to insist on the use of deductive logic, particularly for the sake of obtaining *locally-refutable proofs.* Of course, that "restriction" does not preclude the use of defeasible reasoning (which is easy to do in deductive logic, provided you don't insist on elegant, widely-applicable schema), but rather just makes it stand out, often as the weakest part of an argument. I claim, moreover, that any really solid use of a defeasible reasoning pattern, such as appeal to expertise, can be formulated, with enough effort and perhaps a little creativity, as a really solid deductive argument. Two examples of special classes of deductively-strong appeals to expertise follow.

Example A man, John Doe, is on trial for vehicular homicide. A forensics expert testifies that from his examination of the skid marks on the road and the tires of Doe's car after the accident, he is "certain" (subtext: as certain as he ever is on a judgement like this), that Doe's car was traveling at least 75mph just before the point where the skid marks begin. The proposition that the prosecuting attorney wants to use is a formalization of "John Doe's car was traveling at least 75mph just before it began to skid". In the initial version of the attorney's informally-interpreted proof, she uses a propositional variable X to represent that statement (i.e. the assertion that the statement is true), and another propositional variable Y for a statement that quotes the full record of the testimony of the expert witness, and asserts that it is in fact what the expert said. She then includes the axioms Y and $Y \Rightarrow X$.

Why am I calling this a deductively strong appeal to expertise? It's because the prosecuting attorney, with the help of her expert witness, is quite prepared to replace those two axioms with a longer proof from a larger set of axioms, each of which is much more trustworthy than $Y \Rightarrow X$. Those axioms include axioms about measurements taken by the investigators, which can be checked against crime scene photos and evidence collected at the scene, as well as the physics-based assumptions that the expert uses to derive a lower bound on the speed of the car, e.g. from the length of the skid marks, upper bounds on the force of friction of the tires against the pavement (as a function of the distance along the skid), weight of the car, wind resistance, etc.

Example Frustrated with accusations of biasness and lack of rigor, the climate change experts involved in the Fifth Assessment Report of the UN Intergovernmental Panel on Climate Change take steps to clarify the meaning of their highest-certainty 10-year predictions. From those predictions[21], they formalize a family of increasingly-weak assertions $A_1, \ldots, A_k$, to an

[21] By "predictions", I have in mind conditional statements, possibly with many premises, e.g. "IF at least

extent that independent third parties, after 10 years, can check whether the assertions hold. Together with their government's politicians, they then sign a contract, which stipulates a family of increasingly-lucrative sets of legal entitlements $S_1, \ldots, S_k$ for oil-producing nations (e.g. which permit high levels of pollution), such that S_i is awarded if A_i turns out false. In the aftermath of the IPCC's move, for those highest-certainty predictions, the accusations of biasness and lack of rigor fall off.

This example demonstrates a distinction between two kinds of appeal to expertise. As in the previous example, to formalize their argument, I could start by using X for the prediction of the expert, i.e. one of the assertions A_i. And again we will have axioms X and $Y \Rightarrow X$, for some additional propositional variable Y. The legal entitlements S_i justify our strengthening of $g(Y)$ from some elaboration of "The IPCC experts say that A_i will almost certainly turn out true", to some elaboration of "The IPCC experts say, *and clearly believe*, that A_i will almost certainly turn out true".

At first reading, the difference between the two versions of $g(Y)$ may seem too informal to be meaningful. And indeed, with or without the entitlements, it would be acceptable to make a semantic criticism (defined in Section) about the second version of $g(Y)$. What makes the difference between the two versions of $g(Y)$ meaningful is that, without the entitlements, I do not see the author of the argument being able to adequately formalize "clearly believe" (they will get stuck after a sequence of rigor criticisms and other dialog moves, made with the intention of forcing them to clarify what they mean), whereas with the entitlements it is a simple matter, since the IPCC scientists will also be arguing elsewhere that, *assuming* A_i, a policy should be put into place that would conflict with the legal entitlements S_i (In other words, essentially all that needs to be claimed is that the experts have a strong desire to avoid the granting of the entitlements S_i.).

3.5 Complications/obstacles

3.5.1 Tedium

The proofs in this paper are tedious to read. They were tedious to write, also.

> *The true reason for this straying from the portal of knowledge is, I believe, that principles usually seem dry and not very attractive and are therefore dismissed with a mere taste.*
>
> (Gottfried Leibniz, 1679, *"On the General Characteristic"*[LL76])

In this section I'll discuss why that is, and what, if anything, can be done about it.

3.5.2 Expanding Scope (expanding language)

There is one problem that the dialog system in Section 3.3.2 deals with in (what I consider) an unsatisfactory way, though it is not clear that it can be dealt with much better. General expansions of the language of an argument (which amounts to the scope of the argument)

N tons of oil are burned in the next 10 years AND no major geoengineering project is initiated, then ..."

can be used to block progress indefinitely. On the other hand, we can rarely specify all the relevant concepts before an argument begins. The theoretical dialog system that I will present allows any number of expansions of the language that are accompanied by some other mark of progress, and a pre-specified limit on the number of general expansions (with no mark of progress) that either party can do.

This section will expand on this problem, and discuss whether or not it is practically important for my project.

4 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.[22] In [BHO75] 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.

	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 [BHO75],[23] 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,

[22] 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.

[23] 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.

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/absense 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. $\mathcal{L}_{\mathsf{vague}}$ consists of:

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

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

- 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[24] are as follows. With respect to the definition of *vaguely-interpreted formal proof* from Section 3.2, they are all *assumptions* as opposed to *simplifying assumptions*.

$$\begin{array}{rcl}
\begin{array}{l}\mathsf{App}, \mathsf{Acc}, \mathsf{Acc}_{\mathsf{hyp}}, \mathsf{App}^m, \mathsf{App}^f, \\ \mathsf{App}^1, \ldots, \mathsf{App}^6\end{array} & : & \mathcal{A} \\
|\cdot| & : & \mathcal{A} \to \mathbb{N} \\
0, 1, 512, 825, \ldots & : & \mathbb{N} \\
\cap, \cup & : & \mathcal{A} \times \mathcal{A} \to \mathcal{A} \\
+, -, * & : & \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q} \\
/ & : & \mathbb{Q} \times \mathbb{Q} \to_? \mathbb{Q}^{24} \\
< & : & \mathbb{Q} \times \mathbb{Q} \to \mathbb{B}^{25}
\end{array}$$

[24] Besides =, which is untyped.

4.1 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.

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

$g(\langle \textsf{bias only evidence} \rangle)$ 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."

$g(\langle \textsf{gender uncor with ability in each dept} \rangle)$ we take to be just "Assumption 1" from [BHO75], which I quote here:

> *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.* [BHO75]

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

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

$g(\mathcal{A})$ = "The powerset of $\textsf{App}$. 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.

[24] $\rightarrow_?$ denotes the type of a partial function. The version of many-sorted FOL I use has build-in (AKA "first-class") partial functions.

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

- $<$ 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[26], 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 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 10. For $g \in \{m, f\}$ and $d \in \{1, \ldots, 6\}$:

$$\mathsf{App} := \mathsf{App}^m \uplus \mathsf{App}^f$$

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

Definition 11. 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 $g \in \{m, f\}$:
$\langle\mathsf{gender\ uncor\ with\ ability\ in\ each\ dept}\rangle \Rightarrow$

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

$$|\mathsf{Acc}^{d,g}_{\mathsf{hyp}}| \in \left[|\mathsf{Acc}^{d}_{\mathsf{hyp}}| \cdot \frac{|\mathsf{App}^{d,g}|}{|\mathsf{App}^{d}|} \pm {}^1\!/_2 \right]$$
$$\frac{|\mathsf{Acc}^{d,g}_{\mathsf{hyp}}|}{|\mathsf{Acc}^{d}_{\mathsf{hyp}}|} = \frac{|\mathsf{App}^{d,g}|}{|\mathsf{App}^{d}|}$$

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 $d \in \{1, \ldots, 6\}$:

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

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

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

$$\mathsf{accRate}^{g} := \mathsf{Acc}^{g} / \mathsf{App}^{g} \quad \text{and} \quad \mathsf{accRate}^{g}_{\mathrm{hyp}} := \mathsf{Acc}^{g}_{\mathsf{hyp}} / \mathsf{App}^{g}$$

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

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

then $\langle \textsf{lawsuit should be dismissed} \rangle$

Simplifying Assumption 1. $\langle \textsf{bias only evidence} \rangle$

Claim 1.

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

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 4.3 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.

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

4.2 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 10 and 11 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.[27]
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. $\mathsf{accRate}^f_{\mathrm{hyp}} > \mathsf{accRate}^m_{\mathrm{hyp}}$

[27] 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.

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 $\mathsf{accRate}^{f}_{\text{hyp}} > \mathsf{accRate}^{m}_{\text{hyp}}$ 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.

4.3 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.

5 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/

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.

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 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.

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.

5.1 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 accomodate, 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 α_{min} and α_{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 α_{min} and α_{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 refering 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-constrainted 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 meaningul 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.[28] This is Simplifying Assumption 3. It is probably the most objectionable of the assumptions. I give some criticisms of it in Section 5.3.
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.

[28] 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.

5.2 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 $\{\mathrm{bin}_1, \mathrm{bin}_2, \mathrm{bin}_3, \mathrm{bin}_4\}$.

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 $\mathrm{bin}_1, \mathrm{bin}_2, \mathrm{bin}_3$, 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 $\mathsf{G} = \mathrm{true}$ and $\overline{\mathsf{G}}$ to abbreviate $\mathsf{G} = \mathrm{false}$.

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[29] 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–

[29] 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.

hair fragments that were part of his 2-inch "picky dreads."

H is a simulation random variable whose distribution is the uniform distribution over $\mathsf{Clipped}$, i.e. it is a random hair clipping.

$\mathsf{BParams}$ 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: $\mathsf{BParams}$ 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 is the set of tripples $\langle p_1, p_2, p_3 \rangle \in [0,1]^3$ 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 13. 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.

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

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.[30]

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

width is the function from $\mathsf{Clipped}$ to $\{\text{bin}_1, \text{bin}_2, \text{bin}_3, \text{bin}_4\}$ such that $\text{width}(h)$ 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 indepedently at random with replacement from some hair-width distributions.

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

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

[30] "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.

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.

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

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.

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

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.

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

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 corrolary 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 ⟨the newspaper hair evidence is neutral or exculpatory⟩.[31]

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

Claim 5. If $\alpha_{\min} \leqslant .849$ then $\frac{\mathsf{P(Widths{=}\vec{b}|G)}}{\mathsf{P(Widths{=}\vec{b}|\overline{G})}} < 1$
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 obain a corrolary 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.[32] 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 a man (40,000), and an average quantity of scalp hairs for a man with black hair (110,000).[33] 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. □

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

[33] 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.

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. □

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 $1 - p_1 - p_2 - p_3$.

$$\int_{\vec{p}=\langle p_1,p_2,p_3\rangle \in \mathrm{B}_1} t(p_1, p_2, p_3, 1-p_1-p_2-p_3) d\vec{p} = \int_{p_1=0}^{1/4} \int_{p_4=p_1}^{\frac{1-p_1}{3}} \int_{p_3=p_4}^{\frac{1-p_1-p_4}{2}} t(p_1, 1-p_1-p_3-p_4, p_3, p_4') dp_1 dp_4 dp_3$$

$$\int_{\vec{p}=\langle p_1,p_2,p_3\rangle \in \mathrm{B}_2} t(p_1, p_2, p_3, 1-p_1-p_2-p_3) d\vec{p} = \int_{p_1=0}^{1/4} \int_{p_4=p_1}^{\frac{1-p_1}{3}} \int_{p_2=p_4}^{\frac{1-p_1-p_4}{2}} t(p_1, p_2, 1-p_1-p_2-p_4, p_4) dp_1 dp_4 dp_2$$

$$\int_{\vec{p}=\langle p_1,p_2,p_3\rangle \in \mathrm{B}_3} t(p_1, p_2, p_3, 1-p_1-p_2-p_3) d\vec{p} = \int_{p_1=0}^{1/4} \int_{p_2=p_1}^{\frac{1-p_1}{3}} \int_{p_4=p_2}^{\frac{1-p_1-p_2}{2}} t(p_1, p_2, 1-p_1-p_2-p_4, p_4) dp_1 dp_2 dp_4$$

$$\int_{\vec{p}=\langle p_1,p_2,p_3\rangle \in \mathrm{B}_4} t(p_1, p_2, p_3, 1-p_1-p_2-p_3) d\vec{p} = \int_{p_1=0}^{1/4} \int_{p_2=p_1}^{\frac{1-p_1}{3}} \int_{p_3=p_2}^{\frac{1-p_1-p_2}{2}} t(p_1, p_2, p_3, 1-p_1-p_2-p_3) dp_1 dp_2 dp_3$$

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. $\square$

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. $\square$

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.

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

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

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

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

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} = $ 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. $\square$

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 5.2 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 5.2, 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 5.2 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. $\textit{likelihood-ratio}(.849, 1) < .997$
2. For $\alpha_1 < .849$ have $\textit{likelihood-ratio}(\alpha_1, 1) < \textit{likelihood-ratio}(.849, 1)$
3. For $\alpha_1 < .849$ and $\alpha_1 < \alpha_2 < 1$ have $\textit{likelihood-ratio}(\alpha_1, \alpha_2) < \textit{likelihood-ratio}(\alpha_1, 1)$

5.3 Criticism of argument

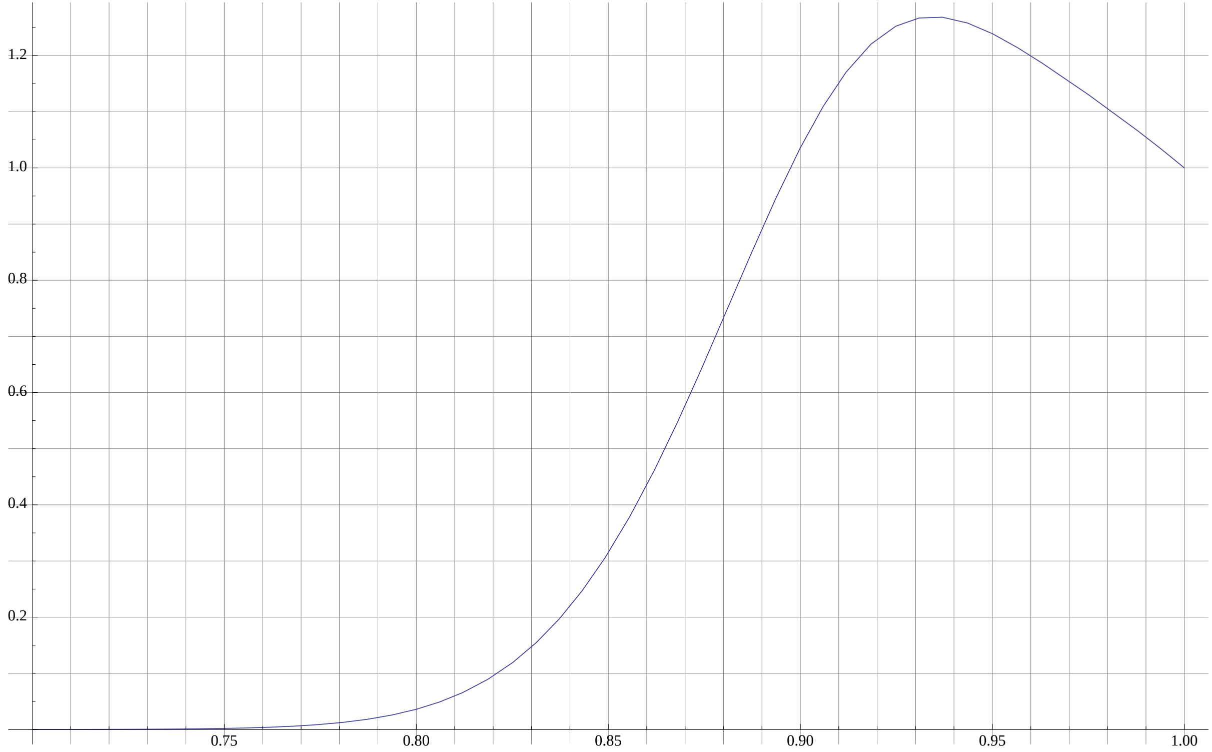
5.3.1 Criticism 1

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.[34] 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 $\mathsf{G} = \text{true}$, and y is the value of the p_1 that maximizes the probability of the evidence given $\mathsf{G} = \text{false}$, then $\int_{\vec{p} \in \mathrm{B}} p_1 \leqslant (x + y)/2$.

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

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

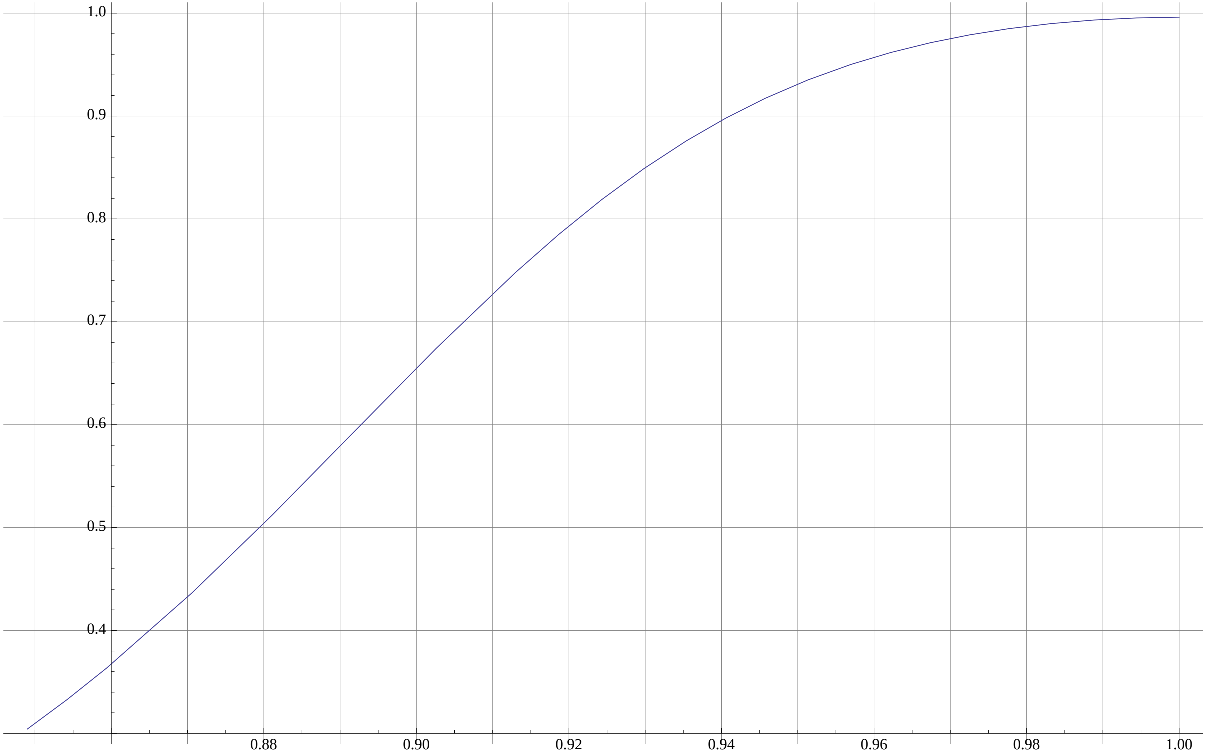


The defense can respond to the criticism, and I will show that. This requires slightly strengthening the hypotheses of Claims 5 and 6.

5.3.2 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 $\{\mathrm{bin}_1, \ldots, \mathrm{bin}_5\}$, where bin_5 is a new constant. The results of that approach are discussed in Section 5.3.3.

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.



5.3.3 Response to criticisms

We can address both criticisms at once; if we introduce a fifth component of $\mathsf{BParams}$ corresponding to the interval $(187.5, \infty)$, and like the first component (probability width is in bin_1) of $\mathsf{BParams}$ 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 $\mathsf{BParams}$ 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.

5.3.4 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 [TCFK83] (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*.[35]

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:

> 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 accomodate 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).

6 Example: The first studies linking smoking and cancer

The following excerpt is from Michael J. Thun's article *When truth is unwelcome: the first reports on smoking and lung cancer*[Thu05]. In my thesis I will compare his quantitative conclusion to mine.

> In retrospect, the strength of the association in the two largest and most influential of these studies – by Enest Wynder & Evarts Graham in the *Journal of the American Medical Association (JAMA)*. . . and by Richard Doll & Austin Bradford Hill (both of whom were later knighted for their work) in the *Briiish Medical Journal*– should have been sufficient to evoke a much stronger and more immediate response than the one that actually occurred. Had the methods for calculating and interpreting odds ratios been available at the time, the British study would have reported a relative risk of 14 in cigarette smokers compared with never-smokers, and the American study a relative risk of nearly 7, too high to be dismissed as bias.

In 1950 two landmark papers were published giving the first strong statistical evidence that tobacco smoking causes cancer, the first in the United States and the second in England. It was not until 1965 that cigarette packages were required to have health warnings in the US. I will give part of an argument here that said policy was well-justified already in the early 1950s. I may flesh out the remainder of the argument later, which involves introducing two more *candidate models* (see below), the cigarette companies' *unknown genotype model*

[35] 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.

and the statistician R.A. Fisher's *soothing herb model*. The easiest way to refute those is to incorporate the data on female smoking and cancer, which neither model is able to explain.[36] The part of the argument given here simply compares a version of the standard, causal model dependModel, to the naive null-hypothesis model indepModel, which posits that smoking and lung cancer are independent. I call dependModel the "dependent-variables" model, since it doesn't actually formalize *why* it predicts that smoking and cancer are dependent variables.

This argument is an instance of the following setup: An experiment to measure some variable is designed and published, with the possible outcomes of the experiment (values of the variable) defined precisely. Sufficient time is given for all the interested parties to publish competing models for predicting the outcome of the experiment, by giving probability distributions over the set of possible outcomes. The experiment is performed. Suppose that one of the models M is "overwhelmingly better" (defined in the experimental design - below, via the definition of Beats$(\cdot,\cdot)$ and Axiom 22) at predicting the true outcome than the others. Moreover, suppose that M asserts that the use of a certain product (may) pose a health risk to its users; below, this is productWarning(M). Then the result of this competition must be communicated to potential users of the product. The warning can be revoked if M loses in a later equally-rigorous experiment competition.

The purpose of this example is, in part, to demonstrate that the requirement of deductive reasoning is not a limitation for problems in the domain I specified (Section 1.1)[37], provided at least that one is firmly committed to certain ideals of persuasion.

6.1 Refinements of the argument

In the argument below, the causal scientific theory, which motivates the assumptions made by dependModel, is not made explicit. With the addition of Fisher's *soothing herb model* and the tobacco companies' *unknown genotype model* (i.e. adding those models to the set AllCM), it would be necessary to make candidate models derive their outcome distributions from other sorts of assumptions. The reason is that those models are contrived to fit the data; they have outcome distributions similar to dependModel's, in order to prevent dependModel from winning on purely quantitative grounds, as it does against indepModel. Hence it is necessary to have a test that at least requires that a model's outcome distribution is derived from some more-readily-understandable axioms. In Fisher's model, the readily-understandable axioms essentially say that lung cancer causes smoking. In the unknown genotype model, they say that there is a common genetic cause of both lung cancer and a person's propensity to smoke tobacco.

[36] Smoking became popular among men years before it became popular among women, and the lung cancer rates reflect this. The unknown genotype model could explain the earlier, smaller rates of lung cancer and smoking among women by suggesting a sex-linked genotype; however, they would not be able to explain why the rates increased so quickly. As for Fisher's *soothing herb model* (lung cancer causes smoking, because of the soothing effect of smoking), it would require an additional hypothesis, unrelated to the purported soothing effect, to explain why there was a delay in the increase of female lung cancer rates.

[37] This example does not today meet the second criteria (contentiousness) that I listed there, but it did in the 1950s.

6.2 Proof with hypergeometric distributions contingent on an unproved mathematical claim

Vaguely-defined sorts (in $\mathcal{L}_{\mathsf{vague}}$)

- CM : candidate models for the possible outcomes of the British study. In the current version of this argument, a candidate model M is determined by outcomeDistr$(M, \cdot)$ and productWarning(M).
- A : set of adult men living in the US at the time when the American study was done.
- B : set of adult men living in England at the time when the British study was done.

Sharply-definable sorts (in $\mathcal{L}_{\mathsf{math}}$)

- $\mathbb{R}$ and $\mathbb{N}$ - reals and natural numbers
- FS$[\alpha]$ - finite subsets of (the interpretation of) the given sort α. This is a function from sorts to sorts.
- Str - strings over the ASCII alphabet
- StudyOutcomes $\leqslant$ FS$[\mathbb{N}]$ - the set $\{620, \ldots, 649\}$. Before the study is done, we don't know how many of the people with lung cancer are smokers, i.e. $|\mathrm{LC}_B^{\mathrm{samp}} \cap \mathrm{S}_B^{\mathrm{samp}}|$ is unknown. The size of that set is smallest when every person without lung cancer is a smoker, and largest when every person with lung cancer is a smoker, so the set of outcomes of the study (the possible sizes of $\mathrm{LC}_B^{\mathrm{samp}} \cap \mathrm{S}_B^{\mathrm{samp}}$) is $\{|\mathrm{S}_B^{\mathrm{samp}}| - |\overline{\mathrm{LC}_B^{\mathrm{samp}}}|, \ldots, |\mathrm{LC}_B^{\mathrm{samp}}|\} = \{620, \ldots, 649\}$.

Function symbols in $\mathcal{L}_{\mathsf{vague}}$

In the following, a person being a "smoker" means that they smoked at least one cigarette per day during the most-recent period when they smoked.

- B^{pop} : FS$[B]$ is a *hypothetical set*; the population that we imagine the British study samples were drawn from.
- $\mathrm{LC}_B^{\mathrm{pop}}$: FS$[B]$ is the set of people in B^{pop} with lung cancer.
- $\overline{\mathrm{LC}}_B^{\mathrm{pop}}$: FS$[B]$ is the set of people in B^{pop} without lung cancer.
- $\mathrm{S}_{B,i}^{\mathrm{pop}}$: FS$[B]$ is indepModel's guess at the set of smokers in B^{pop}.
- $\mathrm{S}_{B,d}^{\mathrm{pop}}$: FS$[B]$ is dependModel's guess at the set of smokers in B^{pop}.
- $A^{\mathrm{samp}}, B^{\mathrm{samp}}$: FS$[A]$ is the sample of patients used in the American (resp. British) study.
- $\mathrm{LC}_A^{\mathrm{samp}}, \mathrm{LC}_B^{\mathrm{samp}}$: FS$[A]$ is the set of people in A^{samp} (resp B^{samp}) who have lung cancer.
- $\mathrm{S}_A^{\mathrm{samp}}, \mathrm{S}_B^{\mathrm{samp}}$: FS$[A]$ is the set of smokers in A^{samp} (resp B^{samp}).
- outcomeDistr$(\cdot, \cdot)$: CM $\times$ StudyOutcomes $\to \mathbb{R}$ is the given candidate model's distribution over StudyOutcomes.
- AllCM : FS[CM] - the set of all candidate models. It should contain a candidate model from every interested party.

Defined function symbols (in $\mathcal{L}_{\mathsf{def}}$)

- StudyOutcomes : FS[$\mathbb{N}$] $:= \{620, \ldots, 649\}$. A copy of the sort StudyOutcomes (see above for definition) that resides in the universe. So StudyOutcomes denotes both (1) a sort, and (2) an element of the universe defined to be the set that is the intended interpretation of (1).
- Constants for the complements of some sets:
  - For each symbol $X \in \{\mathrm{LC}_A^{\mathrm{samp}}, \mathrm{S}_A^{\mathrm{samp}}\}$: $\overline{X} := A^{\mathrm{samp}} \backslash X$
  - For each symbol $X \in \{\mathrm{LC}_B^{\mathrm{samp}}, \mathrm{S}_B^{\mathrm{samp}}\}$: $\overline{X} := B^{\mathrm{samp}} \backslash X$
  - For each symbol $X \in \{\mathrm{LC}_B^{\mathrm{pop}}, \mathrm{S}_{B,d}^{\mathrm{pop}}, \mathrm{S}_{B,i}^{\mathrm{pop}}\}$: $\overline{X} := B^{\mathrm{pop}} \backslash X$
- $\Pr_{x \in U}(x \in V_1 \,|\, x \in V_2) : \mathrm{FS}[\alpha] \times \mathrm{FS}[\alpha] \times \mathrm{FS}[\alpha] \rightarrow_? \mathbb{R} := |V_1 \cap V_2 \cap U| / |V_2 \cap U|$
- For each $k \in \{0, 1, 2\}$:
  $\mathrm{testInterval}_k : \mathrm{FS}[\mathrm{StudyOutcomes}] := \{|\mathrm{S}_B^{\mathrm{samp}} \cap \mathrm{LC}_B^{\mathrm{samp}}| - k, \ldots, |\mathrm{S}_B^{\mathrm{samp}} \cap \mathrm{LC}_B^{\mathrm{samp}}| + k\}$
- For each $k \in \{0, 1, 2\}$:
  $$\mathrm{test}_k(M) : \mathrm{CM} \rightarrow \mathbb{R} \;:=\; \sum_{x=\min(\mathrm{testInterval}_k)}^{x=\max(\mathrm{testInterval}_k)} \mathrm{outcomeDistr}(M, x)$$

Predicate symbols in $\mathcal{L}_{\mathsf{def}}$

- $\mathrm{Beats}(M_1{:}\mathrm{CM}, M_2{:}\mathrm{CM}) \leftrightarrow \bigwedge_{k \in \{0,1,2\}} \mathrm{test}_k(M_1) > 1000 \cdot \mathrm{test}_k(M_2)$. Model M_1 beats model M_2 if it assigns much higher probability to the true outcome $|\mathrm{S}_B^{\mathrm{samp}} \cap \mathrm{LC}_B^{\mathrm{samp}}|$, as well as to the intervals of size 3 and 5 around the true outcome. The interval of size 5 is about 17% of StudyOutcomes, and any larger interval would be biased since the interval of size 5 already contains the maximum of StudyOutcomes.
- $\mathrm{BeatsAll}(M_1{:}\mathrm{CM}) \leftrightarrow \forall M_2{:}\mathrm{CM}.(M_2 \in \mathrm{AllCM} \wedge M_1 \neq M_2) \Rightarrow \mathrm{Beats}(M_1, M_2)$ simply says that M_1 beats all the other models in AllCM.

Function symbols in $\mathcal{L}_{\mathsf{math}}$

- $\{x, \ldots, y\} : \mathbb{N} \times \mathbb{N} \rightarrow \mathrm{FS}[\mathbb{N}]$ is the set of naturals from x to y inclusive, or the empty set if $x > y$.
- $+, \cdot : \{\mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}, \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}\}$ (addition and multiplication)
- $- : \mathbb{N} \times \mathbb{N} \rightarrow_? \mathbb{N}$ is subtraction, but undefined if the result is negative.
- $/ : \mathbb{R} \times \mathbb{R} \rightarrow_? \mathbb{R}$ is division, undefined when the second argument is 0.
- $\Sigma_{x=t_1}^{t_2} t_3(x) : \mathbb{N} \times \mathbb{N} \times (\mathbb{N} \rightarrow \mathbb{R}) \rightarrow \mathbb{R}$ is the usual summation binder symbol. The formal syntax is $\Sigma(t_1, t_2, \lambda x{:}\mathbb{N}.t_3)$.
- $\cap : \mathrm{FS}[\alpha] \times \mathrm{FS}[\alpha] \rightarrow \mathrm{FS}[\alpha]$ is set intersection.
- $\backslash : \mathrm{FS}[\alpha] \times \mathrm{FS}[\alpha] \rightarrow \mathrm{FS}[\alpha]$ is set difference.
- $|\cdot| : \mathrm{FS}[\alpha] \rightarrow \mathbb{N}$ is the size of the given finite subset of (the interpretation of) α.
- $\binom{X}{k} : \mathrm{FS}[\alpha] \times \mathbb{N} \rightarrow \mathrm{FS}[\mathrm{FS}[\alpha]]$ is the set of subsets of X of size k.

- $\min(\cdot), \max(\cdot) : \mathrm{FS}[\mathbb{N}] \to_? \mathbb{N}$ are the minimum and maximum elements of a finite set of naturals. Undefined if the set is empty.
- $\mathrm{hyper}(k, s, N, s') : \mathbb{N} \times \mathbb{N} \times \mathbb{N} \times \mathbb{N} \to_? \mathbb{R}$ is the hypergeometric distribution (in the last argument; the other three arguments are parameters), defined when $s' \leqslant s \leqslant N, s \leqslant k \leqslant N$; if a population of size N has s smokers and $N - s$ nonsmokers, and k people are chosen uniformly at random *without* replacement from the population, then $\mathrm{hyper}(k, s, N, s')$ is the probability that the resulting set contains exactly s' smokers.
- $\mathrm{condHyper}(s_1, s_2, X_1, X_2, s_1') : \mathbb{N} \times \mathbb{N} \times \mathrm{FS}[B] \times \mathrm{FS}[B] \times \mathbb{N} \to_? \mathbb{R}$ is a probability distribution (in the last argument; the other four arguments are parameters), defined when $s_1' \leqslant s_1 \leqslant |\mathrm{S}_B^{\mathrm{samp}}| \leqslant N,\ s_1 \leqslant |X_1|,\ s_2 \leqslant |X_2|$. Suppose we have disjoint sets of people X_1 and X_2 , with X_1 having s_1 smokers and X_2 having s_2 smokers. Uniformly at random we choose size-$|\mathrm{LC}_B^{\mathrm{samp}}|$ subsets X_1' of X_1 and X_2' of X_2. Then $\mathrm{condHyper}(s_1, s_2, X_1, X_2, s_1')$ is the conditional probability that X_1' contains exactly s_1' smokers, given that there are $|\mathrm{S}_B^{\mathrm{samp}}|$ smokers in $X_1' \cup X_2'$.

Simplifying Assumption 7. We would change this to a normal Assumption if we included formalizations of Fisher's and the tobacco companies' models also (see section 6.1 above).

$$\mathrm{AllCM} = \{\mathrm{indepModel}, \mathrm{dependModel}\}$$

Axiom 1. Sizes of sets from the American study.

$$\begin{array}{lll}
|\overline{\mathrm{LC}_A^{\mathrm{samp}}}| & = 780 & \text{patients in sample with conditions other than cancer} \\
|\mathrm{LC}_A^{\mathrm{samp}}| & = 605 & \text{patients in sample with lung cancer} \\
|\overline{\mathrm{S}_A^{\mathrm{samp}}} \cap \overline{\mathrm{LC}_A^{\mathrm{samp}}}| & = 114 & \text{nonsmokers with conditions other than cancer} \\
|\overline{\mathrm{S}_A^{\mathrm{samp}}} \cap \mathrm{LC}_A^{\mathrm{samp}}| & = 8 & \text{nonsmokers with lung cancer}
\end{array}$$

Axiom 2.
productWarning(dependModel) = *"Scientific studies have found a correlation between tobacco smoking and lung cancer that is currently best-explained by the hypothesis that smoking causes an increase in the probability that a person will get lung cancer."*

productWarning(indepModel) = "" (the empty string)

Axiom 3. This gives the sizes of the sample sets, and certain subsets of those sets, from the British study. We evaluate the different models on how well they predict the size of $\mathrm{LC}_B^{\mathrm{samp}} \cap \mathrm{S}_B^{\mathrm{samp}}$, given the sizes of $\mathrm{LC}_B^{\mathrm{samp}}, \overline{\mathrm{LC}_B^{\mathrm{samp}}}$, and $\mathrm{S}_B^{\mathrm{samp}}$. A model predicts the size well if its distribution over StudyOutcomes assigns high probability to $|\mathrm{LC}_B^{\mathrm{samp}} \cap \mathrm{S}_B^{\mathrm{samp}}|$ or some close number; this is formalized in the definition of $\mathrm{Beats}(\cdot, \cdot)$.

$$\begin{array}{ll}
|\mathrm{LC}_B^{\mathrm{samp}}| = |\overline{\mathrm{LC}_B^{\mathrm{samp}}}| & = 649 \\
|\mathrm{S}_B^{\mathrm{samp}}| & = 1269 \\
|\mathrm{LC}_B^{\mathrm{samp}} \cap \mathrm{S}_B^{\mathrm{samp}}| & = 647 \\
|\overline{\mathrm{LC}_B^{\mathrm{samp}}} \cap \mathrm{S}_B^{\mathrm{samp}}| & = 622
\end{array}$$

Simplifying Assumption 8 (dependModel posits a hypergeometric distribution)**.** Note that the values of the four parameters are only bounded by the other axioms, especially Assumptions 17, 18, 19, and 20, with the latter two distinguishing dependModel's distribution from indepModel's.

Still, this and Simplifying Assumption 9 are the worst of the axioms with respect to the standards that I strive for. Unlike the others, we cannot seriously claim that this axiom is literally true with respect to the informal intended semantics, simply because the authors of the British study did not methodically randomize the way that they chose their sample sets of men with and without lung cancer. I would be satisfied to have an axiom that says outcomeDistr(dependModel, ·) is "close enough" to a hypergeometric distribution, but I have not yet investigated suitable ways of formalizing "close enough," and it is not clear that there would be a benefit in pedagogy or cogency that warrants the added complexity.

$$\begin{aligned} & \forall s{:}\text{StudyOutcomes. outcomeDistr}(\text{dependModel}, s) \\ = \; & \text{condHyper}(|\text{LC}_B^{\text{pop}} \cap \text{S}_{B,d}^{\text{pop}}|, |\overline{\text{LC}_B^{\text{pop}}} \cap \text{S}_{B,d}^{\text{pop}}|, \text{LC}_B^{\text{pop}}, \overline{\text{LC}_B^{\text{pop}}}, s) \end{aligned}$$

Simplifying Assumption 9 (indepModel posits a hypergeometric distribution)**.** Note that the values of the four parameters are only constrainted by the other axioms, especially Assumptions 17, 18, and 21, with the Assumption 21 distinguishing indepModel's distribution from dependModel's.

$$\begin{aligned} & \forall s{:}\text{StudyOutcomes. outcomeDistr}(\text{indepModel}, s) \\ = \; & \text{condHyper}(|\text{LC}_B^{\text{pop}} \cap \text{S}_{B,i}^{\text{pop}}|, |\overline{\text{LC}_B^{\text{pop}}} \cap \text{S}_{B,i}^{\text{pop}}|, \text{LC}_B^{\text{pop}}, \overline{\text{LC}_B^{\text{pop}}}, s) \end{aligned}$$

Assumption 17. This is a conservative axiom for dependModel; a figure from the British study says that the rate of lung cancer in men was 10.6 per 100,000 in 1936-1939, and population data for England in 1951 puts the population at about 38.7 million, hence even if the population from which the British sample was drawn is taken to be the entire nation, if we assume about half the population was male, and that the rate at most trippled from 1939 to 1950, then we should only expect about 6100 men with lung cancer.

$$|\text{LC}_B^{\text{pop}}| \leqslant 7000$$

Assumption 18. This is a conservative axiom for dependModel; it says that of the hospital patients from which the British scientists drew their sample, at most 1 in 6 had lung cancer (in reality it would have been significantly lower).

$$|\overline{\text{LC}_B^{\text{pop}}}| \geqslant 5 * |\text{LC}_B^{\text{pop}}|$$

Assumption 19. If we were to define a best-guess version of the dependent-variables model dependModel, we would set the (unknown) left side of the below inequality equal to the (known) right side (and similarly for Assumption 20. However, the evidence is so strongly in favor of dependModel that this much weaker assumption suffices:

$$\frac{\Pr_{x \in B^{\text{pop}}}(x \in \overline{\text{S}_{B,d}^{\text{pop}}} \mid x \in \text{LC}_B^{\text{pop}})}{\Pr_{x \in B^{\text{pop}}}(x \in \overline{\text{S}_{B,d}^{\text{pop}}} \mid x \in \overline{\text{LC}_B^{\text{pop}}})} \leqslant 3 \cdot \frac{\Pr_{x \in A^{\text{samp}}}(x \in \overline{\text{S}_A^{\text{samp}}} \mid x \in \text{LC}_A^{\text{samp}})}{\Pr_{x \in A^{\text{samp}}}(x \in \overline{\text{S}_A^{\text{samp}}} \mid x \in \overline{\text{LC}_A^{\text{samp}}})} \quad ^{38}$$

[38] $= 3(.0132231/.146154) \approx .27142$

Assumption 20. Same comment as in Assumption 19 applies here.

$$\Pr_{x \in B^{\text{pop}}}(x \in \overline{\mathrm{S}^{\text{pop}}_{B,d}} \mid x \in \overline{\mathrm{LC}^{\text{pop}}_B}) \geqslant 1/3 \cdot \Pr_{x \in A^{\text{samp}}}(x \in \overline{\mathrm{S}^{\text{samp}}_A} \mid x \in \overline{\mathrm{LC}^{\text{samp}}_A}) \quad {}^{39}$$

Assumption 21. The independent-variables model simply posits that, in the population from which the British sample was drawn, the fraction of smokers among people with lung cancer is the same as the fraction of smokers among people with illnesses other than lung cancer.

$$\Pr_{x \in B^{\text{pop}}}(x \in \mathrm{S}^{\text{pop}}_{B,i} \mid x \in \overline{\mathrm{LC}^{\text{pop}}_B}) = \Pr_{x \in B^{\text{pop}}}(x \in \mathrm{S}^{\text{pop}}_{B,i} \mid x \in \mathrm{LC}^{\text{pop}}_B)$$

The next assumption states the intended consequence of one model beating all the others.

Assumption 22. $\forall M{:}\text{CM}.\ \text{BeatsAll}(M) \Rightarrow \text{ShouldRequire}(\text{productWarning}(M))$

Claim 13.

$$\text{condHyper}(s_1, s_2, X_1, X_2, s'_1)$$

equals

$$\frac{\text{hyper}(|\mathrm{LC}^{\text{samp}}_B|, s_1, |X_1|, s'_1) \cdot \text{hyper}(|\overline{\mathrm{LC}^{\text{samp}}_B}|, s_1, |X_2|, |\mathrm{S}^{\text{samp}}_B| - s'_1)}{\sum_{x=\min(\text{StudyOutcomes})}^{\max(\text{StudyOutcomes})} \text{hyper}(|\mathrm{LC}^{\text{samp}}_B|, s_1, |X_1|, x) \cdot \text{hyper}(|\overline{\mathrm{LC}^{\text{samp}}_B}|, s_1, |X_2|, |\mathrm{S}^{\text{samp}}_B| - x)}$$

The above axioms, together with some basic mathematical axioms, prove that for any setting of the free parameters $|\mathrm{LC}^{\text{pop}}_B|, |\overline{\mathrm{LC}^{\text{pop}}_B}|, |\mathrm{S}^{\text{pop}}_{B,i} \cap \mathrm{LC}^{\text{pop}}_B|, |\mathrm{S}^{\text{pop}}_{B,d} \cap \mathrm{LC}^{\text{pop}}_B|$, etc that obeys the constraints given by Axioms (17)-(21), the dependent-variables model decisively beats the independent-variables model:

Conjecture 1.

$$\bigwedge_{k \in \{0,1,2\}} \text{test}_k(\text{dependModel}) > 5000 \cdot \text{test}_k(\text{indepModel})$$

From Conjecture 1, the **goal sentence** follows:

$$\text{ShouldRequire}(\text{productWarning}(\text{dependModel}))$$

6.3 Simpler, complete proof

We add a symbol for the binomial distribution.

$$\text{binDistr}_{\cdot,\cdot}(\cdot) : [0,1] \times \mathbb{N} \times \mathbb{N} \to_? \mathbb{N}$$

[39] = (1/3).146154 = .048718

We may either give $\text{binDistr}_{\cdot,\cdot}(\cdot)$ a prose definition, and then state the next axiom as a Claim, or we could make $\text{binDistr}_{\cdot,\cdot}(\cdot)$ a defined function symbol. Either way is consistent with the definition of *vaguely-interpreted formal proof.*

$$\forall p{:}(0,1).\forall n,t{:}\mathbb{N}.\ (0 \leqslant t \leqslant n) \Rightarrow \text{binDistr}_{p,n}(t) = \binom{n}{t} p^t (1-p)^{n-t}$$

We introduce a family of probability distributions that takes the place of $\text{condHyper}(\cdot,\cdot,\cdot,\cdot,\cdot)$. In this case, we give it the following prose definition and state the later two axioms 14 and 15 as Claims, which are made only for the purpose of calculation.

- $\text{condBinom}(p_1, p_2, s'_1) : [0,1] \times [0,1] \times \mathbb{N} \to \mathbb{R}$ is a probability distribution (in the last argument; the other two arguments are parameters). Suppose we sample $|\text{LC}_B^{\text{samp}}|$ times from each of two binomial distribution, the first having success probability p_1 and the second having success probability p_2. Then $\text{condBinom}(p_1, p_2, s'_1)$ is the conditional probability that we get s'_1 successes from the first distribution *given* that the sum of successes is $|\text{S}_B^{\text{samp}}|$.

We also introduce three new constants $p_{\text{S}_d|\text{LC}}, p_{\text{S}_d|\overline{\text{LC}}}$, and $p_{\text{S}_i|*}$ of type $[0,1]$. $p_{\text{S}_d|\text{LC}}$ and $p_{\text{S}_d|\text{LC}}$ are dependModel's estimates of the fraction of smokers in the lung cancer population and in the population of people with conditions other than lung cancer. $p_{\text{S}_i|*}$ is indepModel's estimate of the fraction of smokers in both populations.

We drop Simplifying Assumptions 8 and 9, replacing them with the following two:

Simplifying Assumption 10 (dependModel posits a binomial distribution)**.** Note that the values of the two parameters (first two arguments) of $\text{condBinom}(\cdot,\cdot,\cdot)$ are only bounded by the other axioms, namely Axioms (23), and (24), with the latter two distinguishing dependModel's distribution from indepModel's.

The following paragraph is the same as in the description of Simplifying Assumption 8.

Still, this and Simplifying Assumption (11) are the worst of the axioms with respect to the standards that I strive for. Unlike the others, we cannot seriously claim that this axiom is literally true with respect to the informal intended semantics, simply because the authors of the British study did not methodically randomize the way that they chose their sample sets of men with and without lung cancer. I would be satisfied to have an axiom that says $\text{outcomeDistr}(\text{dependModel}, \cdot)$ is "close enough" to a binomial distribution, but I have not yet investigated suitable ways of formalizing "close enough," and it is not clear that there would be a benefit in pedagogy or cogency that warrants the added complexity.

$$\begin{aligned}&\forall s{:}\text{StudyOutcomes}.\\ &\quad \text{outcomeDistr}(\text{dependModel}, s) = \text{condBinom}(p_{\text{S}_d|\text{LC}}, p_{\text{S}_d|\overline{\text{LC}}}, s)\end{aligned}$$

Simplifying Assumption 11 (indepModel posits a binomial distribution)**.**

$$\begin{aligned}&\forall s{:}\text{StudyOutcomes}.\\ &\quad \text{outcomeDistr}(\text{indepModel}, s) = \text{condBinom}(p_{\text{S}_i|*}, p_{\text{S}_i|*}, s)\end{aligned}$$

The next two axioms bound the frequencies mentioned in the previous two axioms. The description of Axiom (19) in the last section has some motivation that applies here as well.

Assumption 23.

$$\frac{1}{2} \cdot \mathrm{Pr}_{x \in A^{\mathrm{samp}}}(x \in \overline{\mathrm{S}_A^{\mathrm{samp}}} \mid x \in \overline{\mathrm{LC}_A^{\mathrm{samp}}}) \leqslant 1 - p_{\mathrm{S}_d|\overline{\mathrm{LC}}} \leqslant 2 \cdot \mathrm{Pr}_{x \in A^{\mathrm{samp}}}(x \in \overline{\mathrm{S}_A^{\mathrm{samp}}} \mid x \in \overline{\mathrm{LC}_A^{\mathrm{samp}}})$$

Assumption 24.

$$\frac{1}{2} \cdot \mathrm{Pr}_{x \in A^{\mathrm{samp}}}(x \in \overline{\mathrm{S}_A^{\mathrm{samp}}} \mid x \in \mathrm{LC}_A^{\mathrm{samp}}) \leqslant 1 - p_{\mathrm{S}_d|\mathrm{LC}} \leqslant 2 \cdot \mathrm{Pr}_{x \in A^{\mathrm{samp}}}(x \in \overline{\mathrm{S}_A^{\mathrm{samp}}} \mid x \in \mathrm{LC}_A^{\mathrm{samp}})$$

The next two axioms tell us how to compute the distributions

Claim 14. For $n = |\mathrm{LC}_B^{\mathrm{samp}}|$ (and recall $|\mathrm{LC}_B^{\mathrm{samp}}| = |\overline{\mathrm{LC}_B^{\mathrm{samp}}}|$) have

$$\mathrm{condBinom}(p_1, p_2, a)$$

equals

$$\frac{\mathrm{binDistr}_{p_1,n}(a) \cdot \mathrm{binDistr}_{p_2,n}(|\mathrm{S}_B^{\mathrm{samp}}| - a)}{\sum_{x=\min(\mathrm{StudyOutcomes})}^{\max(\mathrm{StudyOutcomes})} \mathrm{binDistr}_{p_1,n}(x) \cdot \mathrm{binDistr}_{p_2,n}(|\mathrm{S}_B^{\mathrm{samp}}| - x)}$$

Claim 15. For $n = |\mathrm{LC}_B^{\mathrm{samp}}|$ (and recall $|\mathrm{LC}_B^{\mathrm{samp}}| = |\overline{\mathrm{LC}_B^{\mathrm{samp}}}|$) have

$$\mathrm{condBinom}(p, p, a) = \frac{\binom{n}{a} \cdot \binom{n}{|\mathrm{S}_B^{\mathrm{samp}}| - a}}{\binom{2n}{|\mathrm{S}_B^{\mathrm{samp}}|}}$$

Lemma 1.

$$\bigwedge_{k \in \{0,1,2\}} \mathit{test}_k(\mathit{dependModel}) > 2000 \cdot \mathit{test}_k(\mathit{indepModel})$$

Proof. –I will give an analytic proof in my thesis.–
The independent variables model has no parameters, so $\mathrm{test}_k(\mathrm{indepModel})$ is a constant for each $k \in \{0, 1, 2\}$.Viewing the 3D plot of $\mathrm{test}_k(\mathrm{dependModel})$ as a function of the parameters $p_{\mathrm{S}_d|\mathrm{LC}}$ and $p_{\mathrm{S}_d|\overline{\mathrm{LC}}}$, it is clear that within the range allowed by Axioms (23) and (24), the function is minimized at one of the corner points; for $\mathrm{test}_0(\mathrm{dependModel})$ and $\mathrm{test}_1(\mathrm{dependModel})$ it is minimized when $p_{\mathrm{S}_d|\mathrm{LC}}$ is maximal and $p_{\mathrm{S}_d|\overline{\mathrm{LC}}}$ is minimal; for $\mathrm{test}_2(\mathrm{dependModel})$ the minimum is at the opposite point, when $p_{\mathrm{S}_d|\mathrm{LC}}$ is minimal and $p_{\mathrm{S}_d|\overline{\mathrm{LC}}}$ is maximal.[40] □

[40] The reason for the switch of locations of the minimum is not very interesting. The maximum likelihood model for the American data slightly overestimates the correlation between smoking and lung cancer in the British data, which can be used to explain the minimum for $\mathrm{test}_0(\mathrm{dependModel})$. However, when we broaden the test interval enough, so that we accept all numbers between 645 and 649 as equally-good predictions of the number of smokers among the lung cancer patients, then all the possible overestimating models (i.e. with means greater than 647) are very close to being as good as the model that maximizes the likelihood of the British data, which results in the worst model being the one that maximally *underestimates* the correlation between smoking and lung cancer.

7 More Major Examples

7.1 Sue Rodriguez at the Supreme Court (in progress)

This argument exists only in HTML right now. In my thesis I will include a LaTeXversion also. Click here for the current draft.

It is an argument for granting the right to assisted suicide to a particular individual, as opposed to an argument for an assisted suicide policy, of roughly the same sort as found in Oregon or Holland, that would provide access to assisted suicide to any Canadian who meets certain requirements (which is the goal of Section 7.2).

Note: "Simplifying Assumption 7", etc, refers to an axiom in the HTML document.

I adopt a narrative where the party criticizing the proof is the supreme court justices who voted to deny Sue Rodriguez's petition for access to physician assisted suicide. Exactly the same argument works for the more-recent case of Gloria Taylor; she was initially granted access to assisted suicide by the British Columbia Supreme Court in 2012, but the decision was overturned in 2013. In Rodriguez's (or Taylor's) particular case, no major party to the argument argued that the government would be doing *her* harm by making assisted suicide legal for her (see Simplifying Assumption 7). Thus, the argument comes down to whether allowing Rodriguez. access to assisted suicide would have a negative effect of some sort (against other people - see Simplifying Assumption 5, which will eventually be a lemma proved from more-obvious assumptions; or abstract principles - see Axiom 4 and Axiom 6) that rivals the negative effect of denying her access.

The main goals of the argument are:

1. To clarify the qualitative cost to Sue Rodriguez of denying her access to assisted suicide.
2. To more-precisely state the position that (1) exceeds any cost incurred if the Supreme Court were to grant her access. Or rather, that no such cost has been presented, and because of that she should have been granted access.

7.2 Assisted suicide policy in Canada (in progress)

There is a more-ambitious, and much more difficult, argument to be made about assisted suicide in Canada, namely that some system should be put into place by which access to assisted suicide would be granted to any Canadian who meets certain (very strict) requirements. This requires, first of all, a specification of such a system, which must be done in significant detail, since it is a delicate matter to prevent, as much as possible, instances of *regrettable uses of legal assisted suicide* (which I define in a sufficiently-precise way already in the Sue Rodriguez argument; see URL in Section 7.1). The hairiest part is that we cannot prove or reasonably claim that there is no possibility of a regrettable assisted suicide, as we can with the Sue Rodriguez case. Hence, it is necessary to have axioms, which will be the weakest part of the argument, which together prove that the negative expected "utility" (but not necessarily formalized in terms of real-valued utilities) from the possibility of a regrettable use of assisted suicide is compensated for by the positive expected utility for the people who gain access to assisted suicide.

7.3 Sesardic's analysis of SIDS case (todo)

I intend to formalize Neven Sesardic's excellent investigative work and Bayesian argument [Ses07] about the famous Sally Clark court case, in which a woman was prosecuted for murdering her two infants, while she claimed that they died of SIDS (sudden infant death syndrome). It is a notable example, as Sesardic, a philosopher, contradicts the hasty conclusions of some prominent statisticians (who themselves famously contradicted the physician who gave expert testimony for the prosecution), essentially by applying the same Bayesian quantitative argument, but with much more care taken in constraining the values of the prior probabilities.

A Proof of completeness for the logic from Section 3.1 [todo]

Perhaps I care more than I should about completeness, but since I've worked out the proof already, why not include it for the subset of readers who care also?

More practically, the proof reduces the logic to normal FOL, so it can be used to implement a theorem prover using existing resolution theorem provers.

References

[BHO75] 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.

[Dun95] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *A* rtificial Intelligence, 77(2):321 – 357, 1995.

[Efr05] Bradley Efron. Bayesians, frequentists, and scientists. *J* ournal of the American Statistical Association, 100(469):1–5, 2005.

[Far93] 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.

[Fef] Solomon Feferman. Is the continuum hypothesis a definite mathematical problem?

[GW12] Thomas F. Gordon and Douglas Walton. A carneades reconstruction of popov v hayashi. *A* rtificial Intelligence and Law, 20:37–56, 2012.

[LCS11] 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.

[LL76] G.W. Leibniz and L.E. Loemker. *P* hilosophical Papers and Letters. Number v. 1 in Synthese Historical Library. D. Reidel Publishing Company, 1976.

[Mil00] Dale Miller. Abstract syntax for variable binders: An overview. In John Lloyd, Veronica Dahl, Ulrich Furbach, Manfred Kerber, Kung-Kiu Lau, Catuscia Palamidessi, LuÃsMoniz Pereira, Yehoshua Sagiv, and PeterJ. Stuckey, editors, *C* omputational Logic – CL 2000, volume 1861 of *L* ecture Notes in Computer Science, pages 239–253. Springer Berlin Heidelberg, 2000.

[Pea09] Judea Pearl. *C* ausality: Models, Reasoning and Inference. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[Pra10] Henry Prakken. An abstract framework for argumentation with structured arguments. *A* rgument & Computation, 1(2):93–124, 2010.

[Ses07] 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.

[Sor13] Roy Sorensen. Vagueness. In Edward N. Zalta, editor, *T* he Stanford Encyclopedia of Philosophy. Winter 2013 edition, 2013.

[TCFK83] 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.

[Thu05] Michael J Thun. When truth is unwelcome: the first reports on smoking and lung cancer. *B* ulletin of the World Health Organization, 83(2):144–145, 2005.

[Wal08] D.N. Walton. *I* nformal Logic: A Pragmatic Approach. Cambridge University Press, 2008.

[Wal11] D.N. Walton. Finding the logic in argumentation. *I* nternational Colloquium, Inside Arguments, 2011.

[WK95] D. Walton and E. Krabbe. *C* ommitment in Dialogue: Basic concepts of interpersonal reasoning. State University of New York Press, Albany NY, 1995.