

RIGOROUS DEDUCTIVE ARGUMENTATION
FOR SOCIALLY RELEVANT ISSUES

by

Dustin Wehr

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

© Copyright 2015 by Dustin Wehr

Abstract

Rigorous Deductive Argumentation
for Socially Relevant Issues

Dustin Wehr
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2015

The most important problems for society are describable only in vague terms, dependent on subjective positions, and missing highly relevant data. This thesis is intended to revive and further develop the view that giving non-trivial, rigorous deductive arguments concerning such problems –*without* eliminating the complications of vagueness, subjectivity, and uncertainty– is, though very difficult, not problematic in principle, does not require the invention of new logics (classical first-order logic will do), and is something that more mathematically-inclined people should be pursuing. The framework of *interpreted formal proofs* is presented for formalizing and criticizing rigorous deductive arguments about vague, subjective, and uncertain issues, and its adequacy is supported largely by a number of major examples. This thesis also documents progress towards a web system for collaboratively authoring and criticizing such arguments, which is the ultimate goal of this project.

Acknowledgements

I have a lot of people to thank. Each group is in alphabetical order.

–My brilliant and open-minded supervisors Professor Stephen Cook and Professor Alasdair Urquhart.

–My parents, Dr. Judith Bloomer and Dr. Robert J Wehr.

–My friends, Lily Bernstein, Kate Busby, Sam Caldwell, Isabel MacKay-Clackett, Mike Markovich, Deborah Perkins-Leitman, Natalie Wiseman

Contents

1	Introduction	vi
1.1	What this project is and isn't	ix
1.2	Role of formal logic	xi
1.3	Preface to examples	xii
1.4	Related Work	xii
1.4.1	Logical Positivism, Logical Empiricism, and Analytic Philosophy Since the Early 1900s	xiii
1.4.2	Mathematics and Theoretical Computer Science	xiv
1.4.3	Informal Logic, Defeasible Reasoning, and Intentional Logics	xiv
2	Proofs and critiques	1
2.1	Problem domain	1
2.2	Interpreted proofs and critiques	2
2.2.1	Criticizing interpreted formal proofs	5
2.3	Choice of logic	7
2.4	Implementation for reading and verifying interpreted formal proofs	14
2.5	Toy example: Walton's fallacious argument demonstrating equivocation via "variability of strictness of standards"	17
2.5.1	Formal criticism of Walton's marriage argument	22
2.5.2	Literal, ungenerous interpretation of (non-simplifying) assumptions	25
3	Classical deductive formalization of defeasible reasoning	27
3.1	Argument from expert opinion	28
3.2	Bayesian reasoning	31
3.3	Theory comparison	35
3.4	Costs/benefits analysis	36
3.5	Counterfactual reasoning, hypothetical scenarios	37
3.6	Multiplicity of reasons	39

4	Example: Sue Rodriguez’s supreme court case	41
5	Example: Berkeley gender bias lawsuit	52
5.1	First argument	54
5.2	Second argument	57
5.3	Data Axioms	59
6	Example: Leighton Hay’s wrongful conviction	60
6.1	High-level argument	61
6.2	Argument	64
6.3	Criticism of argument	73
6.3.1	Criticism 1	73
6.3.2	Criticism 2	74
6.3.3	Response to criticisms	75
6.3.4	An open problem	75
7	Example: Arguing that smoking causes cancer in 1950	77
7.1	Extensions and refinements of the argument	78
7.2	Proof with hypergeometric distributions contingent on an unproved mathematical claim	79
7.3	Simpler, more-easily completable proof	86
8	Example: Assisted suicide should be legalized in Canada	90
9	Ongoing work	108
9.1	Web system for collaborative authoring and criticizing of interpreted formal proofs, and a minimal dialogue system	108
9.1.1	Related work from Informal Logic	108
9.1.2	Design of a web system	110
9.2	Obstacles for this project	112

Chapter 1

Introduction

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

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”[\[LL76\]](#))

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

There are thousands and thousands of pages by hundreds of scholars that are tangentially related to this project; papers about vagueness in the abstract,¹ the theoretical foundations of Bayesian reasoning,² *abstract argumentation 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. This thesis concerns a project of the latter kind; the work is on important specific problems, attacking general theoretical problems only as-necessary. In this way, we avoid getting hung up on details that don't matter.

Surprisingly, it is the (normative side of the) field of Informal Logic that is probably most related to this project [Wal08][WK95]. For a long time now the researchers in that field have understood that dialogue-like interactions, or something similar, are essential for arguing about the problems we are concerned with here (Section 2.1). But 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 they follow in their professional work.³

There are important commonalities between proofs in mathematics and proofs about subjective and vague concepts. For example, in both domains, we only need to axiomatize the structures we are thinking about precisely enough for the proof to go through; our proofs about numbers and sets never⁴ require complete characterizations, and similarly, for proofs about people, laws, moral values, etc, there is no need to fully *eliminate* the vagueness that is inherent in axiomatizations with multiple distinct models. That observation is materialized in this project's use of top-down, minimally-reductionist formal proofs –my name for formal proofs where one does not strive to minimize the number of fundamental (not defined) symbols, or the number of axioms (an assertion remains an assertion until someone demands it become a proved lemma).⁵ I believe top-down, minimally-reductionist formal proofs are the only option when reasoning faithfully about vague concepts.

¹See [Sor13], where the approach I take to reasoning in the presence of vagueness does not appear to be covered. I call my approach vagueness as plurality of intended models.

²I recommend [Pea09].

³[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.

⁴Except for proofs about finite structures.

⁵As a non-essential demonstration of the concept of top-down, minimally-reductionist proofs, and of the dynamic HTML output I've developed for reading arguments, here are two examples that I wrote while debugging my current system. The first is fully formally verified by a first-order theorem prover. Infinitely-many primes: http://www.cs.toronto.edu/~wehr/thesis/infinitely-many_primes.html
5 color theorem: <http://www.cs.toronto.edu/~wehr/thesis/5colorththeorem.html>

There are three aspects of contentious socially-relevant questions that distinguish them from questions that are commonly considered mathematical⁶: vagueness⁷, subjectiveness⁸, and uncertainty. None of these can be eliminated completely without changing the fundamental nature of the problems.

With mathematics problems we can *usually* axiomatize structures sufficiently-precisely at the beginning of our attempt to resolve a problem (a statement to prove or disprove), whereas in reasoning about social issues one must delay the precisifying of vague definitions until necessary – in particular, until critics of one’s argument are too unclear about one’s informal semantics for a symbol to be able to evaluate whether they should accept or reject an axiom that depends on that symbol (this is called a *semantics criticism* in Section 2.2.1). Of course, questions about vague concepts cannot always be answered in a particular way. What may happen is that the question has different answers depending on how it is precisified, which is determined by the author of an argument that purports to answer that question (and sometimes, indirectly, by the critics of the argument). An illustrative example of this can be made with Newcomb’s Paradox⁹; for all of the many English presentation of the problem that I have seen, it is not hard to give two reasonable formalizations that yield opposite answers, a fact that has been ignored, downplayed, or overlooked by many commentators arguing that one of the answers is the *right* answer (and likewise for many puzzles or paradoxes argued about in analytic philosophy).

As with vagueness, subjectiveness demands some system of interaction between people on the two sides of an argument, and I am working on an implementation of such a system now (Section 9.1). Of course I do not mean to suggest that formal logic can help two parties with conflicting beliefs come to the same answer on, say, questions of ethics. However, where formal logic can help is to find *fundamental* sources of disagreement starting from disagreement on some complex question (which is progress!).

Uncertainty is the most difficult of the three complications. Sparsity of information can make it impossible to give an *absolutely*-strong deductive argument for or against a given proposition, and the inability to do so can easily deflate one’s motivation to make a formal demonstration. But interaction is useful here, too: In Chapter 6, I give a proof that a key piece of evidence that was used to convict a man of murder has no inculpatory value. Now, I cannot say that the assumptions from which that conclusion (the proposition named ⟨the newspaper hair evidence is neutral or exculpatory⟩) follows

⁶But note that, as this thesis will make clear, my opinion is that there is no sharp qualitative boundary between the two domains.

⁷Classic examples are vague predicates expressing tallness, or baldness.

⁸E.g. the weights of various principles of morality

⁹Start at the Wikipedia page if you haven’t heard of this and are curious.

are *absolutely* easy to accept, but I confidently challenge anyone to come up with a proof of the negation of that conclusion, i.e. a proof that the likelihood of the convicted man being guilty given the evidence is significantly larger than the likelihood that he is innocent given the evidence. Hence, I am claiming that my assumptions *are* easy to accept *relative* to what my opponents could come up with.

I use a superficial extension of classical FOL in this project, and *for this particular project*—rigorous deductive reasoning about the kind of questions described in Section 2.1—that seems to be the right fit. It is vital that the interface between syntax and semantics is as simple as possible, and classical FOL with Tarski semantics is the best in this respect. In Section 2.3 I make that argument in more detail. In Chapter 3, I take some space to explain how common forms of defeasible reasoning can be carried out in deductive logic, and how ideas from e.g. modal (or, by extension, epistemic or temporal) logic can be used as-necessary without needing to build them into the definition of the logic (complicating the semantics).

1.1 What this project is and isn't

What it is:

This thesis gives the foundations for, and documents progress toward, a collaborative web system intended for arguing about certain kinds of questions in the most rigorous, fair, and civil way that we know of: formal deductive proof. The main proposed uses/benefits of the system are:

1. Making progress in arguments about questions usually considered outside of math and science.
2. Helping people find the fundamental sources of their disagreements with each other (a special case of progress).
3. Demonstrating deductive thinking to people who are not interested in mathematical problems, or are not mathematically inclined.
4. Giving mathematicians an outlet for advocacy work that utilizes their technical abilities in an essential way.¹⁰
5. Serving as a practically-approachable ideal for rigorous, fair, and civil argumentation.

The apparent contradiction between items 3 and 4 is resolved by noting that people who are mathematically inclined and people who are not have different roles in the system. In

¹⁰Of course statisticians have always engaged in advocacy work. By “mathematicians” here, I mean those in disciplines whose overwhelmingly central focus is on proving theorems.

particular, authoring a new argument requires at least one person who is familiar with basic formal logic, but contributing to new and existing arguments, or criticizing arguments, does not. Moreover, as a small number of excellent lawyers and philosophers have demonstrated over the centuries, writing natural language arguments that approximate the ideal (item 5) does not necessarily require familiarity with formal logic. I choose, perhaps over-generously, to interpret Leibniz’s (latest, and most pessimistic) writing describing his imagined *universal characteristic*, with those notes in mind:

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, 1706¹¹)

What it isn’t:

This project is not about using logic for discovery, as in the *axiomatic method*. And it is not concerned with developing general or elegant mathematical theories (including theories presented as “logics”). An important premise of this project’s approach is that an informal yes/no question must be set before formalization begins (though it may later be modified), and only making progress on the question matters (where a more-precise formulation of the question is progress). 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. This preoccupation with constructing relatively-elegant, widely-applicable theories, is one of two factors to which I attribute the lack of success of the Logical Positivists’ project, the other being their focus on questions outside of the Problem Domain (Section 2.1).

This project is fundamentally *normative*. It is not concerned with descriptive modelling of argumentation, as in *abstract argumentation systems* [Pra10]. There is no interest here in modelling real legal reasoning, for example, nor in assisting it. But there is great interest here in depicting what legal reasoning *should* be like in an ideal environment with lots of

¹¹From translation of a letter to Sophia of Hanover [LCS11]

time available (e.g. for severe criminal cases, in which there is the rest of the convicted person's life to argue that they are innocent with sufficiently-large probability¹²).

1.2 Role of formal logic

Remove formal logic from this project and there is no benefit over our current system of arguing with each other through papers and blog posts and shouting. Those mediums are easier to work in, and superior if one is interested in persuasion. Indeed, interpreted formal proofs (the format for arguments and criticisms advocated in this thesis – see Section 2.2) can be made more persuasive to most people if converted to an informal argument that mixes natural language and mathematics in the normal way we use in conference and journal papers. The problem with that, from the point of view of this project, is that unsound, invalid, misleading, unfair, and otherwise bad arguments benefit from the lax regulations as much *or more than* good arguments do. This project uses deductive formal logic because it is our best tool for forcing the weaknesses of arguments to be exposed. Thus **the role of formal logic in this project is regulatory, and nothing more than that.** The success of this project rides on the regulatory benefit outweighing the overhead of formalization.

I have caught omissions in my own reasoning thanks to the constraints of formal deductive logic, things that never occurred to me in thinking and talking about an issue for years, resulting in my having to temper my opinion, sometimes temporarily and sometimes long term. I have gained respect for my opponents on *every* issue that I have attempted arguments about, having been forced to consider all the subtle details (e.g. Canada's lifetime ban against blood donations from men who have had sex with men,¹³ assisted suicide in Canada, the evidence for anthropogenic global warming¹⁴). It is hard

¹²Noting that a convict is adversely effected for the remainder of their life, even if they are released.

¹³The efficacy of our system for preventing contaminated blood from ending up in the stock of blood donations relies not just on tests for HIV, Hepatitis, etc, but also on self-disclosure of known infections and risk-factors. If the ban is lifted, is there non-negligible probability of a significant increase in the rate of people lying in the self-disclosure part of the system? Reasoning deductively, one *must* consider this non-obvious question, and I have found no way to derive my target conclusion (that, with additional safeguards, the ban should be lifted) without making an assumption that is not far from explicitly answering the question 'no'. We could imagine, for example, that after lifting the ban, some homophobic HIV-positive person intentionally donates blood in retaliation. That may seem far-fetched, but it *must* nonetheless be ruled out (with high probability), one way or another. That strictness imposed by formal deductive logic should be reassuring.

¹⁴The closest thing we have to a strong deductive argument, that I have found, comes from the Berkeley Earth Surface Temperature project, which has mostly been ridiculed by climate science researchers, who simply view it as making no significant advancement in climate science, ignoring or not valuing the fact that it seeks to minimize the use of argument from expert opinion.

to write a person off as ignorant or stupid, relative to oneself, after a great struggle to find acceptable formal assumptions from which it follows logically that they are wrong. It is hard to overstate this advantage of rigorous deduction. Among other things it provides a force for compromise – a force in the same democratic spirit as “I could be wrong”, but with much greater discriminatory power.

1.3 Preface to examples

On one hand, the better the quality of the example arguments in Chapters 4-8, the more seriously this project and its theoretical ideas will be taken. There is some basis for that; the expected value of the project and theoretical ideas are much harder to assess, so instead one might choose the heuristic of assessing the author, via the remaining material.

I am urging you to resist that temptation. These are not exemplars of interpreted formal proofs. I am not a genius, nor a real statistician, nor an expert in criminal law or human biology. I am not even an especially good mathematician. But once the web system (Chapter 9.1) is ready, so that the project has as good a chance at gaining traction as possible, I will persuade some such people to collaborate with me on new arguments or to author their own. Much, much better examples are still to come.

This, like almost any thesis, is not intended to be read from start to finish. You are encouraged to skip ahead after reading at least Sections 2.1 and 2.2. The most complete and accessible example is the Sue Rodriguez argument, which should be read in HTML, though a static version is given in Chapter 4. The most complete and accessible example written in L^AT_EX is Chapter 5.

1.4 Related Work

Unfortunately this project does not fit well within any current area of research. On the other hand, it would be impossible without certain firmly established work, especially the fundamentals of classical predicate logic and Bayesian reasoning/statistics. I have chosen to cite related work predominately when it is contextually relevant, throughout this thesis. E.g. in Section 1.1 I briefly talk about the field of Abstract Argumentation Theory, which was not helpful in this project, and in Section 9.1.1 I briefly cover one of the implemented projects that has come out of the Informal Logic community.

1.4.1 Logical Positivism, Logical Empiricism, and Analytic Philosophy Since the Early 1900s

In general in analytic philosophy, there has been an immense amount of writing *about* the hypothetical application of formal logic to vague and subjective issues, but very little of the application itself.

The first scholars to have the requisite technical framework of predicate logic, and to attempt to expand its scope to matters outside of traditional mathematics, were analytic philosophers of the early-to-mid 1900s, especially those associated with the movements of Logical Positivism and Logical Empiricism. Unusually in philosophy, there seems to be a general agreement that they failed. See, for example, *The Heritage of Logical Positivism* [Res85]. Before explaining why they failed, it is important to note that they didn't fail at *this* project. Geoffrey Sayre-McCord writes in *Logical Positivism and the Demise of "Moral Science"*¹⁵:

... most of the Logical Positivists were convinced that moral theory is nonsense. They thought their arguments showed that there really is no such thing as "moral science." Moral language, they maintained, is not used to report facts, rather it is simply a tool used to manipulate the behavior both of ourselves and of others.

That is, most of the positivists shied away from working on questions affected by moral relativism, instead spending their formalization efforts on questions from science.

Considering the few analytic philosophers who made serious attempts to reason in formal logic about specific problems outside of mathematics, 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 2.1. Because there is less 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

¹⁵From the previously-cited compendium.

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.

1.4.2 Mathematics and Theoretical Computer Science

The stigma in mathematics 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 paid to work on it. Nonetheless, disrepute is the current state of things, and has been for many decades.

As a consequence, the most related work in mathematics differs from the work required for this project in one very important respect: It is intensionally reductionist. We find this in Game Theory and Decision Theory, for example. In these fields, real social problems are used to inspire and motivate interesting mathematical problems, but little more than that, aside from extremely rare situations when the simplifying assumptions of models are met, and countless disastrous situations in which the mathematics is used without meeting the simplifying assumptions.

I should mention, however, that these fields are progressing, with the simplifying assumptions for some problems becoming more and more palatable. Perhaps we will see strategy-proof mechanisms employed for kidney exchange programs in the future[AFKP13], for example. Though, I doubt it. I expect that researchers will continue to ignore factors that would make their model too messy and unwieldy, and ignore possible changes to their models that would result in mathematically-uninteresting solutions to the real problem (since then there would be nothing to publish).¹⁶

1.4.3 Informal Logic, Defeasible Reasoning, and Intentional Logics

The idea to adopt an asymmetric dialog system (author vs critics) came from work on argumentation systems in Informal Logic; see [WK95] for one of many sources.

¹⁶Minimizing the number of people who die waiting for kidney transplants, for example. The sophisticated algorithms found in the Algorithmic Game Theory literature implicitly use a model in which, for example, there is no possibility of legislative solutions that mandate participation in a kidney exchange (say, for the hospital to be eligible for federal funding) while criminalizing or penalizing lying (see previous source for details about why hospitals might lie).

Chapter 3 is devoted to explaining the inappropriateness of defeasible logic for this project, and Section 2.3 does the same for intentional logics. All work that I have found on argumentation about subjective and vague questions uses defeasible or intentional logics. The vast majority of it is concerned with the construction of formal systems for hypothetical use (i.e. never seriously applied) in reasoning about subjective and vague questions. Unfortunately, these systems are not useful for the task of writing an isolated deductive argument, as they come with the overhead of their own syntax and semantics, and because, as I hope the examples I provide will convey, one cannot expect to find a formal system of axioms that is completely adequate for one's argument; each argument requires at least a slightly new system in order to be formulated in the most natural way, and formulation in the most natural way is vital for approaching the goal of *locally-refutable proofs*(page 6).

Chapter 2

Proofs and critiques

2.1 Problem domain

Provided the uncertainty involved in a problem is not too great, or that it *is* too great but one side of the argument has the burden of proof, it is my view, from several years of working on this project and a decade of thoughts leading up to it, that the main impediments to rigorous deductive reasoning about socially relevant issues are *a)* conventional mathematical modeling difficulty; *b)* conventional mathematical problem solving difficulty¹; and *c)* tedium². These are strong impediments. For that reason, I think it is worthwhile to describe the questions that I think are best-suited for rigorous deductive reasoning. These are *contentious questions with ample time available*. Typical sources of such problems are public policy and law. Without ample time, it may be detrimental to insist on deductive reasoning; as pointed out in many places, when complete heuristic reasoning and incomplete deductive reasoning are the only options, it is probably best to go with the former. Without contentiousness, there is little motive for employing fallacious reasoning and rhetoric to advance one's position, and this, I think, defeats

¹Two of my proofs, the Leighton Hay argument and the smoking/cancer argument, are currently contingent on the truth of mathematical statements that I cannot easily prove. This is my attitude about such statements: there are mathematicians out there who can easily prove or disprove them, but I think it would be premature to call upon them until proofs of the statements have actually been demanded by critics (called a *mathematics detail criticism* in the paper). In the meantime, I give some empirical evidence of their truth (in this case, numerical evaluation of a complicated integral, without error bounds). Most importantly, there are other, more-subjective axioms of the proof that are much easier targets for criticism. It may even be wise to build in some precedence in the rules for criticizing an interpreted formal proof, whereby under certain conditions (which aren't obvious to me at present) one must accept the axioms that involve vague and subjective concepts before demanding a proof of a purely-mathematical claim (of course, one should always be able to present disproofs).

²This has been the hardest of the three for me to cope with. My hope is that this impediment will be reduced by making the construction of such arguments a collaborative, social process on the web, with an editor having auto-suggestion and other features of modern IDEs (Section 9.1).

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 characteristic³ 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).

2.2 Interpreted proofs and critiques

This section defines an elaboration of 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, which is amenable to criticism, for an author of a proof to describe their intended semantics in the metalanguage.

The definition of *interpreted formal proof* is tailored for classical many-sorted first order logic, but it will be clear that a similar definition can be given for any logic that has a somewhat Tarski-like semantics, including the usual untyped classical first order logic, or fancier versions of many-sorted first-order logic.⁴ A very minor extension of the usual definition of many-sorted first order logic (where sorts must be interpreted as disjoint sets that partition the universe) with easily-eliminable sort operators is used here and in the examples. A language is just a set of symbols, each of which is designated a constant, predicate, function, sort, or sort-operator symbol. A signature is a language together with an assignment of types to the symbols (or, in the case of sort operators, just an assignment of arities).

There are four kinds of formal axioms that appear in an interpreted formal proof:

³Leibniz is referring to the practical system/method that he envisioned, but was unable to devise.

⁴Earlier versions of this thesis included the syntax and semantics of such a fancier logic. That logic is a little more convenient for formalization, but I discarded it because it introduces another barrier to entry for users of the system, and because its reduction to many-sorted first order logic –the language of resolution theorem provers– introduced too many usually-unuseful axioms that drastically slowed down proof search.

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

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

A signature's language has *sort symbols*, which structures must interpret as disjoint subsets of the universe. A language can also have *sort operator symbols*, which are second order function symbols that can only be applied to sorts. In this project sort operators have a 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 $\mathcal{L}_{\text{vague}}$, introduced below).⁵ A signature assigns sorts to its constants, and types to its function and predicate symbols. In this project, types are mostly used as a lightweight way of formally restricting the domain on which the informal semantics of a symbol must be given (by the language interpretation guide). To see why they are beneficial, suppose that we didn't have them, e.g. that we were using normal FOL. For the sake of clarity, we would nonetheless usually need to specify types either informally in the language interpretation guide, or formally as axioms. In the first case, we inflate the entries of the language interpretation guide with text that

⁵For example, if our proof only needs the power set of one mathematical sort S (in $\mathcal{L}_{\text{rigid}}$), then using a sort operator would have little benefit over just introducing another mathematical sort symbol named 2^S . Arguably one cannot say the same if S is a vague sort (in $\mathcal{L}_{\text{vague}}$), since then we would have to introduce 2^S as a vague sort as well, and there is some value for readers in minimizing the number of vague symbols when possible.

rarely needs to be changed as an argument progresses, and that often can be remembered sufficiently after reading it only once. In the second case, we clutter the set of interesting axioms (e.g. the non-obvious and controversial axioms) with uninteresting typing axioms.

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

An interpreted formal proof is given by the following components. Intuitively, $\mathcal{L}_{\text{rigid}}$ is for symbols that should have the same interpretation in all models.

- 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 sentence labeled **goal**. For each sentence label l we write Γ_l for the sentences in Γ labeled l .
- An assignment of one of the sentence labels **assum** or **simp** to each type assignment of Σ . These typing declarations can be viewed as sentences too, and though they will usually be regular assumptions (labeled **assum**), occasionally it's useful to make one a simplifying assumption (labeled **simp**).
- The sentences in Γ_{defn} define the constant, function, and predicate symbols in \mathcal{L}_{def} . Function and predicate symbol definitions have a form like $\forall x_1:S_1 \dots \forall x_k:S_k. f(x_1, \dots, x_k) = t$ where t can be a term or formula (in the latter case, replace $=$ with \leftrightarrow) and the S_i are sorts.
- $\mathcal{L}_{\text{vague}}, \mathcal{L}_{\text{rigid}}, \mathcal{L}_{\text{def}}$ are disjoint languages, $\mathcal{L}_{\text{vague}}$ does not contain any sort-operator symbols,⁶ and \mathcal{L}_{def} contains neither sort nor sort-operator symbols⁷.
- g is a language interpretation guide for a subset of the language of Σ that includes $\mathcal{L}_{\text{vague}}$ and $\mathcal{L}_{\text{rigid}}$. So, giving explicit informal semantics for a defined symbol is optional⁸.
- Optionally for each axiom, a natural language translation of the axiom.⁹

⁶I 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.

⁷Another 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

⁸This may change after experience with criticizing proofs on the web (Section 9.1), as the intended semantics of a defined symbol can be very obscure relative to the intended semantics of the symbols used in the definition, despite the fact that the former semantics is completely determined by the latter.

⁹This was added late to the definition of interpreted formal proofs, as it appears to introduce another source of semantics problems (see "translation criticism" below). The dilemma is that natural language translations will be demanded by readers of interpreted formal proofs as they are usually easier to read, and simply positing that such translations, when given, should not be trusted for veracity, will not make it so. If this becomes a problem, adding features of English to the logic may be useful. On the other

- Γ_{goal} is provable from $\Gamma_{\text{assum}} \cup \Gamma_{\text{simp}} \cup \Gamma_{\text{claim}} \cup \Gamma_{\text{defn}}$.
- For each $\psi \in \Gamma_{\text{claim}}$, any reader in the intended audience of the proof can come up with a set of $\mathcal{L}_{\text{rigid}}$ -sentences Δ , which are true with respect to the (informal) semantics given by g , such that $\Gamma_{\text{assum}} \cup \Gamma_{\text{defn}} \cup \Gamma_{\text{simp}} \cup \Delta$ proves ψ . The first paragraph of the next section gives a more-precise condition.

2.2.1 Criticizing interpreted formal proofs

$\mathcal{L}_{\text{rigid}}$ 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 $\mathcal{L}_{\text{rigid}}$ -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 $\Gamma_{\text{assum}} \cup \Gamma_{\text{defn}} \cup \Gamma_{\text{simp}}$ proves every claim in Γ_{claim} . If that is not the case, then there is some symbol in $\mathcal{L}_{\text{rigid}}$ that should be in $\mathcal{L}_{\text{vague}}$, 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 $\Gamma_{\text{assum}} \cup \Gamma_{\text{simp}}$) independent of the other axioms. When that is not the case, the reader should raise a semantics criticism (defined below), which is analogous to asking “What do you mean by X?”.

In more detail, to review an 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 $\mathcal{L}_{\text{vague}}$ that occurs in ϕ , and report that $g(s)$ is not clear enough for you to *evaluate* ϕ , which means to conclude that all, some, or none of your personal Σ -structures satisfy ϕ . If you cannot resolve this criticism using communication with the author in the metalanguage, then you should submit a Σ -sentence ψ to the author, which is interpreted by the

hand, see Footnote 10 on page 6 about how this is a special case of language interpretation guide entries for defined symbols.

author as the question: Is ψ 's truth compatible with the intended semantics given by g ?

- (2) Classification criticism: Criticize the inclusion of a symbol in $\mathcal{L}_{\text{rigid}}$, if necessary by doing the same as in (1) but for $\mathcal{L}_{\text{rigid}}$. 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 $\mathcal{L}_{\text{rigid}}$ interpretations).
- (4) Subjective criticism: Reject some sentence $\phi \in \Gamma_{\text{assum}} \cup \Gamma_{\text{simp}}$, 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) Strongly reject ϕ : Tentative commitment to $\neg\phi$, i.e. that all of your personal Σ -structures falsify ϕ .
 - (b) Weakly reject ϕ : Tentative commitment to the independence of ϕ , i.e. 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.
- (5) Translation criticism: Criticize as *misleading* the informal natural language text attached to an axiom.¹⁰

In the context of its intended audience, we say that an interpreted formal proof is locally-refutable if no member of the intended audience raises semantic or classification criticisms when reviewing it. A locally-refutable proof has the desirable property that by using the language interpretation guide g , any member of the audience can evaluate each of the axioms of the proof independently of the other axioms. Local-refutability is the ideal for interpreted formal proofs. It is a property that is strongly lacking in most mathematical arguments in economics or game theory, for example, and in every sophisticated abuse of statistics. When an argument is far from locally-refutable, it is hard to criticize in a standard, methodical manner, and that ease of criticism is a central goal of this project.

¹⁰This is actually a special case of criticizing the optional semantic description attached to a definition, since one can always replace an axiom A with a new defined 0-ary predicate symbol $P_A \leftrightarrow A$ and the new axiom P_A .

2.3 Choice of logic

For this project I use a superficial extension of classical many-sorted first-order logic (MSFOL), itself a superficial extension of classical first-order logic (FOL), and *for this particular project* –rigorous deductive reasoning about the kind of questions described in Section 2.1– I believe that is the right choice:

Claim: Nothing simpler than classical FOL will suffice, and nothing significantly more complex is necessary.

Note 1: Though there are many vocal advocates of nonclassical logics, the scope of the previous claim –this project only– is narrow enough that there may be no serious dispute of it, putting it outside of the problem domain described in Section 2.1 due to lack of contentiousness. I will therefore not put the effort into giving a rigorous deductive argument for the claim, and indeed my argument will be neither rigorous nor deductive. That said, I believe that I could find such a formalization without significantly weakening my position in the process.

Note 2: I do believe that the claim holds more broadly –that nonclassical logics should generally be framed and thought of as mathematical theories, as opposed to the grandiose framing as *alternatives to classical logic*¹¹– but it is not relevant to this project to argue that here. Maria Manzano in [Man96] gave arguments in support of this, demonstrating in one book that second-order logic, type theory, modal and dynamic logics can be naturally and usefully simulated in MSFOL. Her book appears to have been mostly ignored by researchers in applied nonclassical logic, and according to [Ven98] she made too strong a claim about the usefulness of translations to MSFOL. I do not advocate that researchers should use the syntax of MSFOL in all situations, since often a custom syntax is more concise, and allows for more-natural statements of metatheorems and simpler programming of automated theorem proving tools (e.g. if one is programming a decision procedure for some version of propositional temporal logic, it would be silly and inefficient to use a problem encoding with bound variables).

Back to the topic of this section: the choice of logic for *this project*. I will focus on the second part of the above Claim, as arguments about the need for the expressiveness of FOL, in particular its predicate symbols and variables, are commonplace.

The main factor compelling the use of FOL is this: It is vital for this project that the interface between syntax and semantics is as simple and transparent as possible, and FOL

¹¹Which has social implications that I won't go into, e.g. giving outsiders the impression that formal logic is controversial or in flux.

with Tarski semantics is the best in this respect. I know of no *substantial* dispute of the relative simplicity and transparency of Tarski semantics. There are criticisms of material implication, of course, but such criticisms stem from misuse of classical logic, in particular use motivated by a desire to extract or impose some sort of meaning from validity that is different from its characterization in terms of truth-with-respect-to-structures. Unfortunately, there are no official directions for the use of FOL, but if there were, they would clearly imply that to reason formally using other sorts of implication, e.g. relevant implication, causality, or conditional probability, one should develop a mathematical theory of relevant implication, causality, or conditional probability, and formalize it as a first-order theory. Since such mathematical theories, without fail, are more complex and nuanced than Tarski semantics, we have no substantial dispute of the relative simplicity and transparency of FOL with Tarski semantics.

The second factor compelling the use of classical FOL, part of which is alluded to in the previous paragraph’s counter-criticism of criticisms of material implication, is this: extensions or similarly-expressive alternatives to FOL are either (1) unuseful or conflicting with the goals of this project, or (2) useful and compatible with the goals of this project, but simulating their benefits for specific proofs is not hard, and so the added complexity of complicating the logic is not justified. This claim is clarified by the following

Hypothesis:

Let L be any nonclassical alternative to MSFOL with its own semantics. I hypothesize that the difficulty of converting a proof in L to an equally-readable proof in MSFOL, or of extending the definition of MSFOL to accommodate the useful features of L ¹², is proportional to the difficulty, compared to MSFOL, of interpreting sets of L -sentences using L ’s semantics. Furthermore, whenever there is no added difficulty of interpreting sets of sentences, or when the features of L allow us to write *easier* to interpret (sets of) sentences,¹³ then I hypothesize we can already simulate those features with low overhead in MSFOL, and/or we can easily extend our definition of MSFOL to accommodate those features, without straying much from Tarski semantics (a “superficial extension”).

I should clarify that the Hypothesis is not an argument against the study of defeasible and nonclassical logics in general, because its force depends on some uncommon aspects of this project:

- Sentences about vague, subjective concepts and uncertain knowledge are by nature already more difficult to interpret than sentences about traditional mathematical

¹²Which I have done, and undone, several times, ending with MSFOL plus type operators.

¹³As I believe is the case for type operators (AKA parametric polymorphism), subtyping, and [Far93]-style partial functions, although the latter two features have drawbacks when it comes to proof search using currently existing theorem provers.

concepts and certain knowledge.

- With the top-down, minimally-reductionist approach to proofs that I advocate, there are fewer syntactic definitions (more non-defined symbols) and more axioms, making ease of interpretation more important than it usually is in applications of formal logic.
- Theory-construction is not a goal; we only care about individual proofs. Because of this, simulation of the used features of another logic L can be *proof specific*, often making the task much easier than a general translation of all L -sentences.

In Chapter 3 I will explain why the arguments used to motivate defeasible logic do not apply to this project, and give some minimal examples of formalizing defeasible reasoning in deductive logic.¹⁴ In contrast to there, where I argue that there is nothing to be gained from using defeasible logic for this project, here I will briefly argue that something would be lost.

Claim 1: The requirements of formal deduction make interpreted formal proof versions of *all* defeasible arguments less psychologically persuasive, though bad defeasible arguments suffer worse than good ones. If defeasible logic were permitted, there would be much less incentive to ever do the extra work required for formal deduction, as bad defeasible arguments can be more psychologically persuasive than good deductive ones.

Claim 2: Argumentation with defeasible logic is not profoundly different from natural language argumentation (e.g. see Section 9.1.1 about the online defeasible argumentation system Carneades), whereas argumentation with deductive logic is. The novelty of interpreted formal proofs about socially-relevant issues is an essential motivation for this thesis, so asking whether defeasible logic should be used reduces to asking whether this thesis should be written at all. Thus, if you are convinced that this thesis was adequately motivated, then you should be convinced that declining defeasible logic was adequately motivated as well.

The remainder of this section, far from considering all alternatives to FOL other than defeasible logics, is devoted to a consideration of modal logics. This is for concreteness, and because, as I said earlier (page 2.3), it is not clear that a deductive argument for

¹⁴The major interpreted formal proofs that take up the bulk of this thesis contain more complex examples, although they are not labeled as such.

why I've chosen FOL is called for. Moreover, *many* of the popular extensions of FOL are modal logics, and modal logics are especially popular in philosophy, which is the field that has historically taken the greatest interest in formal reasoning about contentious issues.

The case against modal logic for this project comes down to two points:

- (i) Modal logic introduces another syntactic device by which semantically-complex sentences can be written in very simple terms. Note that this property is the main *advantage* of modal logic for other applications (other than this project). Note also that FOL (and modal first-order logic, by extension) is not without the capacity to disguise semantically-complex sentences, namely by using deeply nested definitions (but see Footnote 8 on page 4 for discussion of how to address that).
- (ii) Modal logics are easy to simulate in MSFOL, in a natural way, by formalizing possible-world semantics. Hence, in the worst case we would be writing slightly more-verbose sentences (and even that can be mitigated using syntactic definitions). Also, more can be expressed in the translated MSFOL language than can in the language of modal logic.

First let's look at the general simulation, and after that I will use an example to illustrate (i). This is a standard simulation that can be found, in more detail, in [BdRV02].

Let \mathcal{L} be a signature for a many-sorted first-order modal logic with one or more \Box -like modal connectives \Box_1, \dots, \Box_k and with corresponding \Diamond -like connectives \Diamond_j . The corresponding MSFOL signature \mathcal{L}' for the simulation is the same as \mathcal{L} except:

- It has an extra sort \mathcal{W} for worlds. The variables w, w_1, w_2, \dots are reserved for \mathcal{W} .
- For each \Box -like modal connective \Box_j , it has an extra predicate symbol $R_j : \mathcal{W} \times \mathcal{W} \rightarrow \mathbb{B}$ for the corresponding reachability relation.
- Each function or predicate symbol f in \mathcal{L} whose domain type is $S_1 \times \dots \times S_n$ has domain type $\mathcal{W} \times S_1 \times \dots \times S_n$ in \mathcal{L}' . That is, an extra argument is added for the world with respect to which the predicate or function is being evaluated.

Then \mathcal{L} -sentences are easily translated to \mathcal{L}' -sentences by the following function $\ulcorner \cdot \urcorner$, which given an \mathcal{L} -formula A produces an \mathcal{L}' formula with the same number and sorts of free variables as A , plus exactly one free variable w of sort \mathcal{W} . The syntax $t[w \mapsto w']$ means to substitute variable w' for free occurrences of variable w in t .

- $\ulcorner P(t_1, \dots, t_n \urcorner = P(w, \ulcorner t_1 \urcorner, \dots, \ulcorner t_n \urcorner)$ for P a predicate symbol, including $=$.
- $\ulcorner f(t_1, \dots, t_n \urcorner = f(w, \ulcorner t_1 \urcorner, \dots, \ulcorner t_n \urcorner)$ for f a function symbol.
- $\ulcorner A \Rightarrow B \urcorner = \ulcorner A \urcorner \Rightarrow \ulcorner B \urcorner$ and similarly for the other boolean connectives.

- $\lceil \Box_j A \rceil = \forall w_i. R_j(w_i, w) \Rightarrow \lceil A \rceil[w \mapsto w_i]$, where w_i does not occur in A . Thus, $\Box_j A$ holds in the current world w iff A holds in every world that is R_j -reachable from w . Note that in most presentations of frame semantics, the order of the arguments to the reachability relation R is swapped. I use this order so that we can display the translated formula more neatly as $\forall w_i R_j w. \lceil A \rceil[w \mapsto w_i]$.
- $\lceil \Diamond_j A \rceil = \exists w_i. R_j(w_i, w) \wedge \lceil A \rceil[w \mapsto w_i]$, where w_i does not occur in A . We can display such a formula more neatly as $\exists w_i R_j w. \lceil A \rceil[w \mapsto w_i]$.

For any modal logic L that can be characterized by frame semantics, I claim there is a recursive (and usually finite) set of \mathcal{L}' -sentences Ψ_L that capture validity for L . That is, such that for any \mathcal{L} -sentence B and set of \mathcal{L} -sentences A_1, A_2, \dots , have

$$A_1, A_2, \dots \models_L B \quad \text{iff} \quad \Psi_L, \forall w. \lceil A_1 \rceil, \forall w. \lceil A_2 \rceil, \dots \models \forall w. \lceil B \rceil$$

where \models_L is entailment for L and \models is entailment for MSFOL.

Note that, according to the definition of interpreted formal proof, language interpretation guide entries must be given for the worlds sort \mathcal{W} and the reachability relations R_i . Those components of the modal logic semantics are usually *not* given explicit intended interpretations in applications of modal logic in philosophy, and I claim this is the *only* reason why modal logic “paradoxes” in philosophy do not get resolved; the intended semantics remains too vague.

As an example, I will use Fitch’s Paradox of Knowability. It is a proof in the language of a propositional modal logic with either one \Box -like modal connective K or two \Box -like modal connectives K and \Box . I will use the latter formulation, since it is easy to convert to the former by identifying K and \Box . In either case, only one \Diamond -like connective is used, and it is connected to \Box . The axioms are substitution instances of the following axiom schema, where ϕ, ϕ_1, ϕ_2 are metavariables that range over formulas, except for schema (\Box -Valid) where they range over provable formulas. In every presentation of the proof that I have seen, an (overly-simple) English translation of each sentence is provided. I am giving the ones from [CC14], except for ($K\Diamond$ -connection) for which I give two versions.

(K -Factivity)	$K\phi \Rightarrow \phi$	“If a proposition is known, then it is true.”
($K \wedge$ -Distributivity)	$K(\phi_1 \wedge \phi_2) \Rightarrow K\phi_1 \wedge K\phi_2$	“If a conjunction is known, then its conjuncts are also known.”
($\Box \Diamond$ -Connection)	$\Box(\neg\phi) \Rightarrow (\neg\Diamond\phi)$	One direction of “ $\Box\neg\phi$ is logically equivalent to $\neg\Diamond\phi$.” The SEP article[BS13] translates it as “necessarily $\neg\phi$ entails that ϕ is impossible.”
(Knowability Principle)	$\phi \Rightarrow \Diamond K\phi$	“Every truth is knowable.”
(\Box -Valid)	$\Box\phi$, if ϕ is provable from K -Factivity and $K \wedge$ -Distributivity	

From those axiom schema and the rules of classical propositional logic, it follows that for any formula ϕ ,

$$(G) \quad \phi \Rightarrow K\phi \quad \text{“Every truth is known” (the unwanted conclusion)}$$

I will argue that the proof is circular when K is **S5**-like¹⁵, whereas in weaker modal logics G simply does not mean “Every truth is known,” and the meanings of the (Knowability Principle) and (K -Factivity) are far from clear (see second-to-last paragraph of this section for why they are unclear). My position is that the ongoing philosophical discussion about this proof is completely reliant on underspecified semantics.

Since the translation I gave is for first-order modal logic, I’ll briefly describe how to convert to that form. Any nonlogical language can be used, since the proof is really a meta-level proof, but it is convenient to use a minimal language with a single 0-ary predicate symbol P , in which case we can take the goal sentence G to be $P \Rightarrow KP$. Then, the substitution instances of the above axiom schema needed to derive G and justify the

¹⁵More accurately, when the corresponding reachability relation R_K is trivial in that every world is reachable from every other.

instance of (\Box -Valid)¹⁶ are:

(K -Factivity)	$K\neg KP \Rightarrow \neg KP$
($K \wedge$ -Distributivity)	$K(P \wedge \neg KP) \Rightarrow KP \wedge K\neg KP$
($\Box \diamond$ -Connection)	$\Box(\neg K\neg G) \Rightarrow (\neg \diamond K\neg G)$
(Knowability Principle)	$\neg G \Rightarrow \diamond K\neg G$
(\Box -Valid)	$\Box\neg K\neg G$

And the universal closures of the translations of those axioms into MSFOL prove $\forall w. \ulcorner G \urcorner$. In fact, it is easy to see that the universal closures of the translations of ($K \wedge$ -Distributivity) and ($\Box \diamond$ -Connection) are theorems of MSFOL, and (\Box -Valid) can be made a lemma provable from the other axioms, since it is a logical consequence of the (Knowability Principle) and ($\Box \diamond$ -Connection)¹⁷. Therefore, I will give the translations only for the remaining two axioms, since they are the only ones that can be criticized. Note that $\ulcorner G \urcorner = P(w) \Rightarrow \forall w' R_K w. P(w')$.

1. $\forall w. [\forall w_1 R_K w. \neg \forall w_2 R_K w_1. P(w_2)] \Rightarrow \neg \forall w' R_K w. P(w')$, i.e.
 $\forall w. [\forall w_1 R_K w. \exists w_2 R_K w_1. \neg P(w_2)] \Rightarrow \exists w' R_K w. \neg P(w')$. Suppose that for every world w_1 that is K -reachable from the current world, there exists a world w_2 that is K -reachable from w_1 in which P is false. Then there exists a world K -reachable from the current world in which P is false.
2. $\forall w. \neg \ulcorner G \urcorner \Rightarrow \exists w_1 R_{\Box} w. \forall w_2 R_K w_1. \neg \ulcorner G \urcorner [w \mapsto w_2]$

For the class of models that satisfy $\forall w_1, w_2. R_K(w_1, w_2)$, those axioms are equivalent to

1. $\neg \forall w. P(w) \Rightarrow \neg \forall w. P(w)$, a tautology.
2. $\forall w. \neg \ulcorner G \urcorner \Rightarrow (\text{NI}_{\Box}(w) \wedge \forall w_2. \neg \ulcorner G \urcorner [w \mapsto w_2])$ where $\text{NI}_{\Box}(w)$ ('**n**ot **i**solated') abbreviates $\exists w_1. R_{\Box}(w_1, w)$. For the models under consideration, $\forall w_2. \neg \ulcorner G \urcorner [w \mapsto w_2]$ is false, so the consequent in the implication is false, and the sentence further reduces to $\forall w. \ulcorner G \urcorner$.

So, for that class of models the (Knowability Principle) instance is equivalent to G , i.e. the argument is circular.

Thus, we can assume that at least some of the author's intended models of the axioms falsify $\forall w_1, w_2. R_K(w_1, w_2)$. But then we can no longer collapse quantifiers; the two axioms

¹⁶That is, the third and fourth axioms prove $\neg K\neg G$, which is the prerequisite needed to use $\Box\neg K\neg G$ as an axiom.

¹⁷Using the fact that if $\forall w. \phi(w)$ is valid for any MSFOL-formula ϕ , then $\forall w. \forall w' R_{\Box} w. \phi(w)$ is valid also.

1. $\forall w. [\forall w_1 R_K w. \exists w_2 R_K w_1. \neg P(w_2)] \Rightarrow \exists w' R_K w. \neg P(w')$.
2. $\forall w. \neg \ulcorner G \urcorner \Rightarrow \exists w_1 R_{\square} w. \forall w_2 R_K w_1. \neg \ulcorner G \urcorner [w \mapsto w_2]$

cannot obviously be simplified. And without having more-precise semantics for R_K , R_{\square} , and \mathcal{W} , I claim one cannot evaluate whether those axioms are satisfied for the author’s intended interpretations. Now let’s see why it would be hard to give any interesting interpretation of those axioms. The (K -Factivity) axiom schema $K\phi \Rightarrow \phi$ seems acceptable according to the naive interpretation of K when written in the language of modal logic. However, for instances when ϕ itself contains modal connectives, the naive interpretation of the sentence “If ϕ is known, then it is true” is plainly wrong. The reason is that ϕ talks about different objects of the universe (namely, different worlds) when evaluated at different worlds; the sentence should be read as “If ϕ is known, then a certain formula related to ϕ is true”. The same goes for the (Knowability Principle) $\phi \Rightarrow \diamond K\phi$; it is fine to interpret it as “Every truth is knowable” only if ϕ contains no modal connectives. Otherwise, one must say the more verbose “If ϕ is true, then a certain formula related to ϕ is knowable.” This point is similar to the criticism given by Kvanvig in a number of papers beginning with [Kva95] (and later an entire book!).

Of course we did not *need* to translate to MSFOL to make the criticism of the previous two paragraphs. We could have just used the language of frame semantics. But that is missing the point. The criticism explained why the axioms seemed reasonable, which is a requirement for a strong rebuttal in philosophy, and I have demonstrated that it is not hard to do this in MSFOL. But my position –the position of the system advocated in this thesis– is that this is asking too much of a critic. My work in the criticism should have ended much earlier, at the point just before considering the two cases of whether or not there are intended interpretations that falsify $\forall w_1, w_2. R_K(w_1, w_2)$. At that point, I should simply make a semantics criticism with either the (K -Factivity) or (Knowability Principle) instance and one of the symbols R_K , R_{\square} , or \mathcal{W} . The virtue of insisting on writing the axioms in MSFOL is just this: it forces the *author* of the proof to reveal the complexity of their axioms, rather than putting that burden on the critic.

2.4 Implementation for reading and verifying interpreted formal proofs

A good approximation of the format in which I intend interpreted formal proofs to be read –in an effort to make reading them less effortful and tedious– can be seen in any of these examples:

1. Sue Rodriguez's case at the Supreme Court of Canada: http://www.cs.toronto.edu/~wehr/thesis/sue_rodriguez.html
2. Assisted suicide should be legalized in Canada: http://www.cs.toronto.edu/~wehr/thesis/assisted_suicide_msfol.html
3. Walton's intentionally-fallacious argument that no one should get married: http://www.cs.toronto.edu/~wehr/thesis/walton_marriage.html
4. Infinitely-many primes: http://www.cs.toronto.edu/~wehr/thesis/infinitely-many_primes.html
5. 5 color theorem: <http://www.cs.toronto.edu/~wehr/thesis/5colortheorem.html>
6. High-level proof of Gödel's Second Incompleteness Theorem: <http://www.cs.toronto.edu/~wehr/thesis/G2b.html>

An interpreted formal proof is implemented as an HTML document with the following structure:

- A sequence of declarations, each of which is a new symbol introduction (with definitions being a special case), axiom, or lemma. An axiom is either an Assumption, Simplifying Assumption, Assertion (intended to be uncontroversial), Claim (author is a prepared to prove, once challenged), or Quasi-definition (a symbol introduction together with an axiom that is not technically a syntactic definition, but plays a definition-like role).
- The statement of a lemma or theorem A , the *goal* of the interpreted formal proof, which uses only the previously-introduced symbols.
- A collapsible proof of A , which is another interpreted formal proof whose immediate-child declarations, combined with those that preceded the statement of A , entail A , where the entailment is verified by a first-order theorem prover¹⁸ (see below for more detail). Note that this is slightly atypical in that one may delay introducing axioms and primitive symbols until just before they are used in a proof. The purpose of this is to lessen the effect of an interpreted formal proof starting with an overwhelming number of symbol introductions and axioms before it even gets to the statement of the goal sentence. Instead, the declarations that must precede the statement of a lemma are just the symbol introductions for the symbols that are explicitly used in the lemma.

¹⁸“preceded” means the declaration is in its scope, where scope is defined as in many programming languages. For example, if a lemma A_2 immediately follows a lemma A_1 , then the proof of A_2 can use A_1 but nothing introduced in the proof of A_1 , and the proof of A_1 cannot use A_2 .

- Each lemma is the goal of an interpreted formal proof or else it has an informal natural language proof.

Such HTML documents will be a central part of the web system described in Section 9.1. The main advantages over reading in L^AT_EX/PDF are these:

- Collapsible sections of text (implemented). This is helpful as an author 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 a proof/justification of some lemma/claim, or once they have sufficiently-internalized the syntactic definition or informal semantic description of a symbol.
- Pop-up references on cursor hover (implemented). Hovering the mouse cursor over occurrences of symbols will reveal information from their initial declaration. This is more useful for proofs about socially-relevant issues than it usually would be for proofs in mathematics, because of the much higher ratio of

$$\frac{\text{number of fundamental symbols with no standard meaning}}{\text{length of proof}}$$

- Reader comments (implemented): As a temporary stand-in for the plans of Section 9.1, readers can attach annotations (e.g. for criticisms) to any part of an argument (via AnnotateIt.org), or post (nested) comments at the bottom of the page (via DISQUS.com).
- Renameable symbols (in the works). If the name chosen by the author is not conducive to your reading, then change it!
- Multiline display of formulas (in the works). As in some programming languages, there will be a standard format, in terms of where white space is placed, for displaying the structure of formulas across multiple lines, so that an author need only indicate with a checkbox whether the children of a subterm should be displayed on different lines.

The displayed syntax need not be the same as the input syntax. In particular, two distinct symbols can display the same way. Hovering the cursor over an occurrence will reveal which version it is. This takes care of most of the use cases for overloading (where e.g. a function symbol can have multiple function types) while remaining in standard many-sorted first-order logic.

Most of the examples listed above are written in a formal language¹⁹, which gets translated to HTML and to instances of theorem proving problems in many-sorted first

¹⁹The exceptions are examples 5 and 6, which I wrote before implementing the formal language.

order logic, specifically the TFF (typed first-order formula) language of TPTP[Sut09]. Those problems can be solved automatically by first-order theorem provers, although when there are a very large number of axioms and definitions, as in the Assisted Suicide argument (Chapter 8), it is sometimes necessary to tell the prover which axioms to use for each lemma.²⁰ I used SNARK[SWC00] for type checking and sometimes short proofs, CVC4[BCD⁺11] for model finding and sometimes proof search, and Vampire[KV13] for fast proof search and sometimes countermodels, all via the System on TPTP web interface²¹ (except Vampire was also easy to setup locally).

2.5 Toy example: Walton’s fallacious argument demonstrating equivocation via “variability of strictness of standards”

This example is also available for reading in HTML:

http://www.cs.toronto.edu/~wehr/thesis/walton_marriage.html

Here is the informal argument, verbatim from [Wal08]:

1. Getting married involves promising to live with a person for the rest of your life.
2. Nobody can safely predict compatibility with another person for life.
3. One should not make a promise unless one can safely predict that one will keep it.
4. If two people aren’t compatible, they can’t live together.
5. One should not promise to do something one cannot do.
6. Therefore, nobody should ever get married.

Lines 1-4 of the informal argument correspond to Assumptions 1-3 below. Line 5 is redundant, and line 6 corresponds to the proved conclusion, Proposition 1 below.

Sorts:

- P is for people.
- D for dates (e.g. 1 Sept 1998).
- A for potential actions that are associated with a particular date, but not a particular person (like verbs).

²⁰This is due to the non-goal-directed nature of the saturation-based first-order theorem provers that I have used; it is possible that a backwards theorem prover, perhaps even a cut-free proof search, would work better in such cases, but I have not yet found a good, easy to set up implementation.

²¹<http://www.cs.miami.edu/~tptp/cgi-bin/SystemOnTPTP>

- Ψ contains a subset of the ⟨formula, object assignment⟩ pairs (see Definitions 1 and 2). The intended interpretation of this symbol would be simpler if we introduced constant symbols for each element of P and D , in which case I could just identify Ψ with a particular finite set of sentences (of the form of the formulas on the right side of \leftrightarrow in Definitions 1 and 2).

Predicate symbols:

$$\begin{aligned}
 \text{Does} & : P \times A \rightarrow \mathbb{B} \\
 \text{LiveWith} & : P \times P \times D \times D \rightarrow \mathbb{B} \\
 \text{Holds} & : \Psi \rightarrow \mathbb{B} \\
 \text{Compatible} & : P \times P \times D \times D \rightarrow \mathbb{B} \\
 \text{CanSafelyPredict} & : P \times \Psi \times D \rightarrow \mathbb{B} \\
 \text{ShouldNotDo} & : P \times A \rightarrow \mathbb{B} \\
 \leq & : D \times D \rightarrow \mathbb{B}
 \end{aligned}$$

Function symbols:

$$\begin{aligned}
 \text{getMarriedTo} & : P \times D \rightarrow A \\
 \text{makePromise} & : \Psi \times D \rightarrow A \\
 \text{liveWithTillDeath} & : P \times P \times D \rightarrow \Psi \\
 \text{compatibleTillDeath} & : P \times P \times D \rightarrow \Psi \\
 \text{dateOfDeath} & : P \rightarrow D \\
 \text{min} & : D \times D \rightarrow D
 \end{aligned}$$

Style notes:

- The following variables are reserved for the following types: d for D , a for A , p and q for P , and ψ for Ψ . Similarly for the primed and subscripted versions of those variables.
- I leave out leading universal quantifiers.
- To improve readability, when a function symbol takes arguments of type P , I put the arguments in the subscript, as in $\text{Does}_p(a)$, and when a function symbol takes one or more arguments of type D , I put them in the superscript, as in $\text{LiveWith}_{p,q}^d d'$.

Formalization notes;

- It is not hard to correct the argument for the objection that it clearly doesn't work when p and q are near death, in which case it's especially reasonable to reject Assumption 1. I haven't done so since the argument has other, more-serious flaws.

- If after reading the next bullet list about the informal semantics, you would, like me, still reject Assumption 2 for being too broad, then move the sentence to the position of a premise of the goal sentence (Equation 1).
- In retrospect, it would have been more-economical to make $\text{liveWithTillDeath}_{p,q}^d$ and $\text{compatibleTillDeath}_{p,q}^d$ primitive instead of $\text{LiveWith}_{p,q}^{d,d'}$ and $\text{Compatible}_{p,q}^{d,d'}$, but the way I've done it is more faithful to Walton's presentation.

Here is a sketch of part of the informal semantics (i.e. a language interpretation guide):

- $\text{LiveWith}_{p,q}^{d,d'}$ iff p and q are both alive and share the same main residence during the period from d to d' .
- $\text{makePromise}^d(\psi)$ means to make an utterance, on date d , like “I promise that A ”, which is directed at someone with the intention of their interpreting it as a sincere and literal statement.
- The semantics of $\text{compatibleTillDeath}_{p,q}^d$ and $\text{liveWithTillDeath}_{p,q}^d$ are essentially determined by the semantics of the other symbols by Definitions 1 and 2.
- The informal semantics for the other symbols (except for Compatible and CanSafelyPredict , which correspond to the terms in the informal argument that are used with varying “strictness of standards”) are not surprising and not hard to flesh out.

Definition 1. The year when the first of p or q dies.

$$\text{firstDeath}_{p,q} := \min(\text{dateOfDeath}_p, \text{dateOfDeath}_q)$$

Quasi-Definition 1. For all p, q, d there is a proposition $\text{liveWithTillDeath}_{p,q}^d : \Psi$ that holds iff p and q live together from the date d until one of them dies.

$$\forall p, q, d. \text{Holds}(\text{liveWithTillDeath}_{p,q}^d) \leftrightarrow \text{LiveWith}_{p,q}^{d, \text{firstDeath}_{p,q}}$$

Quasi-Definition 2. For all p, q, d there is a proposition $\text{compatibleTillDeath}_{p,q}^d : \Psi$ that holds iff p and q are compatible from the date d until one of them dies.

$$\forall p, q, d. \text{Holds}(\text{compatibleTillDeath}_{p,q}^d) \leftrightarrow \text{Compatible}_{p,q}^{d, \text{firstDeath}_{p,q}}$$

Assertion 1. If p marries q (on date d), then p makes the promise (on date d) that they will live together until one of them dies.

$$\text{Does}_p(\text{getMarriedTo}_q^d) \Rightarrow \text{Does}_p(\text{makePromise}^d(\text{liveWithTillDeath}_{p,q}^d))$$

Assumption 1. Roughly: You can't safely predict that two people will be compatible till death. More precisely: No person p' , on any date, can safely predict that two people p and q will be compatible from that date until one of their deaths.

$$\neg \text{CanSafelyPredict}_{p'}^d(\text{compatibleTillDeath}_{p,q}^d)$$

Assumption 2. If p cannot (on date d) safely predict that ϕ will be true, then p should not (on date d) promise that ϕ will be true.²²

$$\neg \text{CanSafelyPredict}_p^d(\psi) \Rightarrow \text{ShouldNotDo}_p(\text{makePromise}^d(\psi))$$

Assumption 3. Two people who are incompatible during a period cannot live together during that period.

$$\neg \text{Compatible}_{p,q}^{d,d'} \Rightarrow \neg \text{LiveWith}_{p,q}^{d,d'}$$

Assertion 2. If p (on date d) can safely predict that ψ will hold, and ψ implies ψ' , then p (on date d) can safely predict that ψ' will hold.

$$\text{CanSafelyPredict}_p^d(\psi) \wedge (\text{Holds}(\psi) \Rightarrow \text{Holds}(\psi')) \Rightarrow \text{CanSafelyPredict}_p^d(\psi')$$

Assertion 3. If p should not do action a , and doing a' requires doing a , then p should not do a' .

$$\text{ShouldNotDo}_p(a) \wedge (\text{Does}_p(a') \Rightarrow \text{Does}_p(a)) \Rightarrow \text{ShouldNotDo}_p(a')$$

The axioms prove the following, which is the goal sentence:

Proposition 1. $\text{ShouldNotDo}_p(\text{getMarriedTo}_q^d)$

Proof. Let p, q, d be arbitrary. Assertions 1 and 3 imply

$$\text{ShouldNotDo}_p(\text{makePromise}^d(\text{liveWithTillDeath}_{p,q}^d)) \Rightarrow \text{ShouldNotDo}_p(\text{getMarriedTo}_q^d)$$

Hence it suffices to prove

$$\text{ShouldNotDo}_p(\text{makePromise}^d(\text{liveWithTillDeath}_{p,q}^d))$$

²²There is a sense in which it would be more technically correct to write “is true” at both places where I wrote “will be true”, since the truth value of a sentence does not depend on time, but on the other hand “will be true” is consistent with common usage of English, where one can say “I think A will be true” in order to convey the meaning “I think A is true, but we won't know for sure until some point in the future.”

An instance of Assumption 2 is:

$$\neg \text{CanSafelyPredict}_p^d(\text{liveWithTillDeath}_{p,q}^d) \Rightarrow \text{ShouldNotDo}_p(\text{makePromise}^d(\text{liveWithTillDeath}_{p,q}^d))$$

And so it suffices to prove

$$\neg \text{CanSafelyPredict}_p^d(\text{liveWithTillDeath}_{p,q}^d) \quad (2.1)$$

An instance of Assumption 1 gives:

$$\neg \text{CanSafelyPredict}_p^d(\text{compatibleTillDeath}_{p,q}^d) \quad (2.2)$$

Using Definitions 1, 2 and Axiom 3 we can derive:

$$\text{Holds}(\text{liveWithTillDeath}_{p,q}^d) \Rightarrow \text{Holds}(\text{compatibleTillDeath}_{p,q}^d) \quad (2.3)$$

Finally, from (2.2), (2.3) and Assertion 2, (2.1) follows. \square

Assume that the issues I mentioned above under ‘‘Formalization notes’’ have been dealt with. Provided the audience takes the marriage vows seriously (e.g. imagine they are all devout Catholics), you should be able to fill in the language interpretation guide, starting from the sketch I gave above, in such a way that it would be hard for any audience member to reject any of the axioms except for exactly one of Assumptions 1 or 3, and possibly Assumption 2.

Let’s suppose that we accept Assumption 2. So we focus on Assumptions 1 and 3:

$$\text{Assumption 1: } \neg \text{CanSafelyPredict}_p^d(\text{compatibleTillDeath}_{p,q}^d)$$

$$\text{Assumption 3: } \neg \text{Compatible}_{p,q}^{d,d'} \Rightarrow \neg \text{LiveWith}_{p,q}^{d,d'}$$

Furthermore, suppose that I raise a semantics criticism against the symbol CanSafelyPredict , and that our dialogue rules allow me, as a critic, to suggest an extension of the language and axioms. I suggest the introduction of a predicate symbol for personal probability assessment on a particular date,²³ together with the new sort symbol $[0, 1]$ for the real interval $[0, 1]$ ²⁴:

$$\text{Prob} : P \times \Psi \times D \times [0, 1]$$

²³Similar to Bayesian probability, with the intended semantics given in terms of betting games, but without the convention that a probability is assigned to every proposition.

²⁴Or the rationals between $[0, 1]$, or even a finite set such as $\{0, .01, .02, \dots, .98, .99, 1\}$ would suffice.

The intended semantics is for $\text{Prob}_p^d(\phi, x)$ to mean that on date d , person p thinks ϕ is or will be true with probability at least x . And I suggest the axiom:²⁵

$$\text{CanSafelyPredict}_p^d(\phi) \Rightarrow \text{Prob}_p^d(\phi, .8)$$

With that in place, it would be harder for the author of the proof to equivocate about the meaning of CanSafelyPredict . Now, the only symbol whose intended semantics is too vague (whose definition is too incomplete) for us to evaluate Assumptions 1 and 3 is $\text{Compatible}_{p,q}^{d,d'}$. And that brings us to the serious flaw in the informal argument. For Assumption 1 to be true under a given interpretation, the semantic definition of $\text{Compatible}_{p,q}^{d,d'}$ needs to be fairly strong, but for Assumption 3 to be true under a given interpretation, the semantic definition of $\text{Compatible}_{p,q}^{d,d'}$ needs to be quite weak (meaning its extension is large). For example, I would not be abusing the dialog rules if I rejected Assumption 3 for any definition of $\text{Compatible}_{p,q}^{d,d'}$ that is much stronger than this:

p and q are compatible during $[d, d']$ unless one of them poses a physical danger to the other, or one of them makes an effective legal action to remove the other from the household.

And, for such a weak definition of $\text{Compatible}_{p,q}^{d,d'}$, I would have no difficulty justifying my rejection of Assumption 1.

2.5.1 Formal criticism of Walton's marriage argument

Recall from Section 2.5 that $\text{compatibleTillDeath}$ is defined in terms of Compatible and some symbols whose descriptions are clear. The three disputable assumptions were:

Assumption 4. No person p' , on any date, can safely predict that two people p and q will be compatible from that date until one of their deaths.

$$\neg \text{CanSafelyPredict}_{p'}^d(\text{compatibleTillDeath}_{p,q}^d)$$

Assumption 5. If p cannot (on date d) safely predict that ϕ will be true, then p should not (on date d) make a promise that ϕ will be true.

$$\neg \text{CanSafelyPredict}_p^d(\psi) \Rightarrow \text{ShouldNotDo}_p(\text{makePromise}^d(\psi))$$

²⁵If you change .8 to a value much closer to 1, then Assumption 2 becomes easy to dispute.

Assumption 6. Two people who are incompatible during a period cannot live together during that period.

$$\neg \text{Compatible}_{p,q}^{d,d'} \Rightarrow \neg \text{LiveWith}_{p,q}^{d,d'}$$

The initial semantic description (language interpretation guide entry) entry for `CanSafelyPredict` (which is empty, so the only hint we had for interpreting the symbol was the name of the symbol) is too vague for me to evaluate Assumption 5. Specifically, how much confidence must p have in ψ 's truth in order to “safely predict” ψ ? We can say the same for Assumption 4 with either symbol `Compatible` or `CanSafelyPredict`, or for Assumption 6 with symbol `Compatible`, although we should prefer Axioms 5 or 6 since each depends on only one too-vague symbol.²⁶ But supposing I choose Assumption 5, then by the definition of criticizing an interpreted formal proof (Section 2.2) there seems to be only one productive thing to do, which is to make the semantics criticism $\langle \text{Assumption 5}, \text{CanSafelyPredict} \rangle$. I will then communicate with the author directly, suggesting they change the semantic description of `CanSafelyPredict` to something like “If p can safely predict ψ on a given date d (i.e. $\text{CanSafelyPredict}_p^\psi(d)$), then on that date p has credence at least X that ψ is or will be true,” where X is some fixed constant.

Of course the author may reject that suggestion, and instead, for example, add prose to the semantic description of `CanSafelyPredict` that, being still too vague, does not actually help me interpret `CanSafelyPredict` well enough to evaluate Assumption 5. In that case, I will introduce some new symbols, which are under my control, along with an axiom A , also under my control, that uses the new and old symbols and expresses the above Bayesian interpretation of `CanSafelyPredict`. The author can then accept (unlikely, given the previous failure using informal communication), weakly reject, or strongly reject A . This formalizes our disagreement about the meaning of `CanSafelyPredict`, and documents it for later readers of the argument. Suppose the author accepts my suggestion, say for $X = .9$. Then I can accept Assumption 5.

The author’s semantic description of `Compatible` is still too vague for me to evaluate the other two axioms. Once again, according to the definition of criticizing an interpreted formal proof, it seems the only productive thing for me to do is make the semantics criticism $\langle \text{Assumption 6}, \text{Compatible} \rangle$ or $\langle \text{Assumption 4}, \text{Compatible} \rangle$.

As concluded in Section 2.5, if the author clarifies the semantics of `Compatible` and it is very weak (its extension is large), then I can make a *subjective criticism* of Assumption

²⁶This can be useful if informal communication fails, since one can more-easily use the author’s acceptance of the axiom to deduce constraints on the meaning of the too-vague symbol. In particular, one can sometimes, for the sake of making a criticism, simplify an axiom by partially evaluating it using the parts of the author’s language interpretation guide that *are* sufficiently precise.

4 and *strongly reject* it²⁷, and if the author clarifies the semantics of Compatible and it is (at least) moderately-strong, then I can make a subjective criticism of Assumption 6 and strongly reject it. Finally, if the author makes his intended semantics for Compatible somewhere between “very weak” and “moderately strong”, then I can reject *both* of Assumptions 4 and 6. Any of those three scenarios would be good places to end the dialogue.

More likely (in this scenario with such an uncooperative author), the author would see the vulnerability, and avoid clarifying the semantics of Compatible enough that I can make a semantics criticism. In that case, I would formalize the idea of the previous paragraph in the following way. I introduce predicates

$$\begin{aligned} \text{GetAlongOk} & : P \times P \times D \times D \rightarrow \mathbb{B} \\ \text{Murderous} & : P \times P \times D \times D \rightarrow \mathbb{B} \end{aligned}$$

My language interpretation guide entry for $\text{GetAlongOk}_{p,q}^{d_1,d_2}$ says that p and q get along OK during the period $[d_1, d_2]$, and the entry for $\text{Murderous}_{p,q}^{d_1,d_2}$ says that p and q will try to kill each other if they come into contact during $[d_1, d_2]$. Additionally, I introduce two defined 0-ary predicate symbols in order to give names to two sentences:

$$\begin{aligned} \text{CompatVeryWeak} & := \forall p, q, d_1, d_2. \neg \text{Murderous}_{p,q}^{d_1,d_2} \Rightarrow \text{Compatible}_{p,q}^{d_1,d_2} \\ \text{CompatModeratelyStrong} & := \forall p, q, d_1, d_2. \text{Compatible}_{p,q}^{d_1,d_2} \Rightarrow \text{GetAlongOk}_{p,q}^{d_1,d_2} \end{aligned}$$

Finally, I introduce the following axioms (which I accept), which formally describe my above stated positions on the author’s two remaining controversial assumptions for a range of possible precisifications of Compatible (since the author has not made his intended semantics precise). These axioms imply that I reject at least one of those two assumptions:

$$\neg \text{CompatVeryWeak} \vee \neg \text{CompatModeratelyStrong}$$

$$\text{CompatVeryWeak} \Rightarrow (\text{Assump 6} \wedge \neg(\text{Assump 4}))$$

$$\text{CompatModeratelyStrong} \Rightarrow (\text{Assump 4} \wedge \neg(\text{Assump 6}))$$

$$(\neg \text{CompatModeratelyStrong} \wedge \neg \text{CompatVeryWeak}) \Rightarrow (\neg(\text{Assump 4}) \wedge \neg(\text{Assump 6}))$$

²⁷Meaning all of my personal interpretations of the language falsify the axiom. Note that only requires that all of my intended interpretations have at least one tuple $\langle p', p, q, d \rangle$ for which the formula is false.

2.5.2 Literal, ungenerous interpretation of (non-simplifying) assumptions

There is another problematic axiom in argument, Assumption 6, that is easily fixable and which according to the directions for criticizing an interpreted formal proof, *should* be criticized even if the critic knows it is fixable. One simple acceptable way for the author to respond is by changing the label of the axiom from Assumption to Simplifying Assumption.

Instances of Assumption 6 for which d and d' are close should be weakly rejected. Even for a weak definition of compatible (but not quite as weak as `CompatVeryWeak`), there would exist two people who are to that extent strongly incompatible and yet manage to live together for a few days. The author should address the criticism, and here are two quick ways of doing so according to the rules:

1. Make Assumption 6 a Simplifying Assumption, and amend the natural language text associated with it to describe the sense in which it is a simplifying assumption.
2. Add a hypothesis such as $d + 365 \leq d'$ to Assumption 6 (introducing sort \mathbb{N} and symbols $+: D \times \mathbb{N} \rightarrow D$ and $365: \mathbb{N}$ ²⁸), yielding:

$$\left[d + 365 \leq d' \wedge \neg \text{Compatible}_{p,q}^{d,d'} \right] \Rightarrow \neg \text{LiveWith}_{p,q}^{d,d'}$$

With option 2, we can also use the new symbols to formalize an assumption that has the same purpose as the informal constraint on the sort P for people that says P includes only people who are not near death (given in its semantic description in Section 2.5), as follows:

Simplifying Assumption 1. When two people get married, they both live for at least a year after.

$$\text{Does}_p(\text{getMarriedTo}_q^d) \Rightarrow d + 365 \leq \text{firstDeath}_{p,q}$$

Now, what was the point of this pedantry? Essentially, it is the application of a safety principle. The method of formal deduction and criticism with interpreted formal proofs that I advocate may not be robust unless a critic can insist on having technical problems fixed without having to justify why it is important to do so. Otherwise, disputes about meaningful matters will sometimes devolve into unending arguments about argumentation itself. In Leibniz's words:

²⁸365 is larger than necessary, but it is not important.

...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?"

(Gottfried Leibniz, 1679, "On the General Characteristic"[[LL76](#)])

Chapter 3

Classical deductive formalization of defeasible reasoning

The purpose of this chapter is twofold. First, to share the high-level ideas of some reusable formalization patterns that I have used in the course of writing examples. Second, as an extension of Section 2.3 that addresses, by example, objections against the foundations of this project along the lines of deduction being inappropriate for real-world reasoning. The following quote from [Gor88] is an example of such an objection. I have inserted numbers (n) for the purpose of commenting following the quote.

Standard propositional and predicate logics are monotonic. That is, if a proposition is logically implied by some set of propositions, then it is also implied by every superset of the initial set. (0) Another way of describing monotonicity is to say that once something is determined to be true, it remains true. (1) No additional information can cause conclusions to be modified or withdrawn. (2) There is no way to presume something to be the case until there is information to the contrary. (3) There are no rules of thumb, or general rules, which allow conclusions to be drawn which may be faulty, but are nonetheless better than indecision. (4) Classical logic offers no theory about when to prefer one belief to another in general, and provides no language for stating which beliefs to prefer given that certain things are known in a particular case.

(5) The subject matter of classical logic is truth, not decision making. The central concern of logic is logical consequence: which propositions are necessarily true given that other propositions are true. (6) Monotonic logic is very useful when we want to know what must be the case if something else is

known to be true. It is less useful when we know very little about some domain with certainty, or can discover the facts only by extending resources, if at all. (7) Monotonic logic alone provides us with an infinite number of conditional statements of the form "this would be true if that is true", which is of little help in making decisions when we are unable to establish with certainty the truth or falsity of the alternative premises.[Gor88]

Already at (0) we have a classic indicator of problems to come: reference to unqualified, non-relative truth is often not meaningful in formal logic, and definitely is not meaningful for classical FOL, which is defined in terms of truth-with-respect-to-structures. This ambiguous use of “truth” leads to all sorts of confusion and equivocation, and should be banished whenever one is debating the merits of one logic over another.

Points (1)-(4) are a straw man argument. The author implicitly conflates the *use of classical logic via first-order theories* with the *definition of classical FOL*; the former is the author’s desired target, and the latter is the straw man. The fact is that none of these complaints about the literal, technical definition of classical predicate logic apply to *first-order theories*, which are what the author should really be attacking.

Points (5)-(7) are an innocent, understandable, and common oversimplification that I believe leads to an unfortunate misconception about the scope of classical predicate logic, or even formal logic in general, especially among people unfamiliar with formal logic. (5): It would be more accurate to say that the subject matter of classical predicate logic is *semantics*, with relative-truth being an important special case. Perhaps this is easier seen in the formulations of MSFOL that treat the set of truth values as just another sort, so that the boolean connectives are just very common function symbols. (6): It is true that when there is a great deal of uncertainty in a domain, those boolean function symbols are used a little less, with function symbols for Bayesian reasoning having a larger role, but that is hardly a criticism of classical logic. (7): In fact conditionals remain just as essential in domains with a lot of uncertainty. They are our main tool for excluding from consideration the structures that we are not interested in reasoning about, and they are just as useful when those structures contain e.g. Bayesian distributions, interpretations of defeasible legal statutes, etc.

3.1 Argument from expert opinion

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 (lack of) practical effects 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].

That quote is an unusually respectable one of its kind, due to Walton’s explicitly saying “everyday conversational argumentation” and “Basing it on an absolutely universal generalization”. But it still resembles a kind of straw man argument that is often used in the motivation for defeasible logic: implicitly the proponent suggests that in a deductive logic framework, to formalize an argument from expert opinion, 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 (the problem being that no such schema exists). I say that it is slightly more dignified because it is sometimes an innocent instance of academics’ often-very-productive instinct to generalize, and 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* (Section 2.2.1 page 6). 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 0-ary predicate symbol X to represent that statement (i.e. the assertion that the statement is true), and another 0-ary predicate symbol 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 is 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 bias 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 extrapolations. From those extrapolations¹, they formalize a family of increasingly-weak assertions A_1, \dots, A_k , in enough detail 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, \dots, 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 extrapolations, the accusations of bias 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 extrapolation of the expert, i.e. one of the assertions A_i . And again we will have axioms

¹By “extrapolations”, I have in mind conditional statements, possibly with many premises, e.g. “IF at least N tons of oil are burned in the next 10 years AND no major geoengineering project is initiated, then ...”

Y and $Y \Rightarrow X$, for some additional 0-ary predicate symbol Y that says something along the lines that the IPCC experts said A_i will almost certainly turn out true. 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 semantics criticism (page 5) 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 governments who employ the experts have a strong desire to avoid the granting of the entitlements S_i .

3.2 Bayesian reasoning

This section is about how to criticize arguments that use Bayesian reasoning. I use the phrase “subjective probability assumptions” to refer to the informal class of assumptions that includes priors, bounds on conditional probabilities, and choices of parametric models.

Bayesian reasoning and statistics feature prominently in several of my major examples. The literature on the problem of interpreting Bayesian subjective probability assumptions, and the (sometimes insubstantial) Bayesian vs Frequentist debate, is vast (see [F+11] and [Efr05] for refreshingly concrete and pragmatic perspectives). Without surveying all the motivations and interest in the interpretation problem, let us make more precise why it is a problem for this project that cannot be easily dismissed.

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 generally the kind of 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. If you say that your prior for a murder suspect’s guilt is x , what recourse do I have if I think that prior is unreasonable? And how do I make precise what it is that I am disagreeing with? Interpretations involving betting ratios/dutch books [Tal13] help with the latter problem, but not obviously with the former.

I claim that the origin of this uneasiness is the same as that caused by using explicit real-valued utility functions to reason about people’s subjective values. I handle those cases in the same way. The issue in both cases is with the mixture of qualitative and quantitative subjective assumptions.

To be more concrete, let’s look at an example of a typical Bayesian probability assumption, from a famous case in England in which a criminal defense team was allowed to have a statistics expert present a Bayesian analysis to a jury - one of only a few times this has ever been allowed in a jury case [Don05][Kad08]. Most of the details of the case are not important for our purposes; a woman was assaulted by a man, the assailant left DNA evidence, and years later a man was prosecuted, and ultimately convicted, after a “cold hit”² on London’s DNA database made him the suspect. Please note that we won’t be focusing on the DNA aspect of the case, which is the most interesting and contentious part (see [SPMS09], [DF99]; later on I will release an interpreted formal proof about this case). The only reason I mention the DNA aspect is for point 1 on page 34. The probability assumption we are focusing on is the instantiation (or bounding, as I would prefer) of two of the defense team’s fundamental parameters involved in estimating the likelihood ratio:

$$\frac{\Pr(\text{evidence} \mid \text{suspect guilty})}{\Pr(\text{evidence} \mid \text{suspect innocent})^3}$$

Namely these two:

$$c_1 = \Pr(\text{victim failed to identify suspect in police lineup} \mid \text{suspect guilty})$$

$$c_2 = \Pr(\text{victim failed to identify suspect in police lineup} \mid \text{suspect innocent})$$

Now suppose I make assumptions that constrain the parameters of the argument, including those two, enough to imply that the likelihood ratio is large, suggesting –contrary to my opponent’s intuition, let’s assume– that the suspect is guilty. One contributor to that numeric result is that I make an assumption that implies c_2/c_1 is upperbounded by a particular number close to 1, so that the failure to identify the suspect cannot contribute much to making the likelihood ratio large. My opponent will want to reject that assumption, believing that the ratio should be much larger than 1.

Moreover, suppose my opponent believes that my assumption about c_2/c_1 is *unreasonable*. Since these are probabilities representing my credence about whether certain unrepeatable events happened in the past, it is hard for my opponent to argue that I am

²Meaning the man’s DNA was run through the DNA database before he was a suspect.

³Here “innocent” means *factually innocent*, as opposed to *legally not guilty*.

being unreasonable in an objective sense (this is the crux of the Bayesian interpretation problem). Some Bayesians will answer that my assumption is *not* unreasonable provided it is consistent with my other probability assumptions and standard probability axioms. Ultimately I agree with that position, and I offer no magic solution that would enable us to *resolve* situations of fundamental uncertainty. However, in the context of this project, there is something more I can do to at least *state* my position in a clear way.

Before demonstrating my recommendation, I'll briefly remind the reader how the clarification of a position is usually done in Bayesian reasoning. One introduces additional random variables representing evidence or environmental factors, with their own (usually brief) informal interpretations. Then one introduces new independence and conditional probability assumptions to derive a bound on the new likelihood ratio (new because now the meaning of *evidence* has changed). This is indeed the right thing to do when one can precisely describe the domains of the new random variables, and when one has some reason to be confident in the new probability assumptions. But those conditions are often hard to meet, and when they are not met, the additions may introduce more noise and obfuscation into the argument than they do clarification. Consider, for example, trying to formalize a model of the victim forgetting the look of the assailant over time (over a year passed between the crime and the lineup), which is complicated by the dependence on both the look of the (unknown!) assailant and the victim's general ability to recall faces.

My recommendation involves adding much more detail to the informal semantics, only adding detail to the formal mathematical model once agreement has been established for the informal semantics. Suppose I formulate my assumption about the ratio c_2/c_1 as follows. I make two assumptions:

- A1: We can model all our knowledge that is significantly relevant to c_2/c_1 with the following thought experiment, which consists of two scenarios. First, I specify large, precise sets of white men M and women W . In scenario 1, random members m of M and w of W appear in the exact locations of the assailant and victim in the real crime, under similar lighting and weather conditions. Time resumes and a crime may or may not happen, depending on m and w . Any instances of the thought experiment in which events do not transpire in a way sufficiently similar to how they did with the actual crime are ignored. w is shown a police lineup containing m (after a time delay equal to the one in reality), where the other men in the lineup are selected from $M - m$ in a fixed manner (which I would explain in detail) that is similar to how they were selected in reality (roughly based on looking similar to m). Then c_1 is the probability that w fails to identify m . Scenario 2 is the same except that two distinct random members m, m' of M are sampled, with m' representing

an innocent suspect, and w is shown a lineup containing m' instead of m . Then c_2 is the probability that w fails to identify m' .

A2: The second assumption says that c_2/c_1 is upperbounded by a particular number close to 1.

We cannot perform such an experiment, since doing so would be unethical and impractical. Although that verbose pair of assumptions does not get us significantly closer to an objective *test*, it does get us significantly closer to an objective *description*. The motivation is to be more precise, and a few observations about the thought experiment will convey that:

1. The innocent suspect and the assailant are sampled almost-independently of each other (except for the constraint that they are distinct). This is actually a significant modeling assumption! Because, even if we assume the suspect Adams is innocent, we know his DNA test profile is the same as the assailant's – and it is *plausible* that such genetic similarity makes two men significantly more likely to look significantly similar to each other.⁴
2. Assuming W and M are large and varied, I do not use any of the (mostly noisy) knowledge that we have of the victim and suspect. This amounts to an implicit assumption that such knowledge is irrelevant.
3. I have fixed a method of constructing police lineups. In my opponent's favour, I have made no assumption that the lineup was flawed in a way that would make the victim less likely to select a guilty suspect and/or more likely to select an innocent suspect.
4. We cannot test my second assumption A2, but A2 *does* have the virtue of being a relatively-objective assumption (relative to how precisely I specify the thought experiment). I claim this is valuable.

If the thought experiment is actually carried out, then this reasoning strategy is essentially *Empirical Bayes*[Efr05]⁵. Note the structural/qualitative character of assumption A1. It amounts to my modeling assumptions. In contrast A2 has a numeric character. This fits my general recommendation: split a subjective probability assumption into two parts: (A1) a subjective, qualitative part, and (A2) an objective, quantitative part. Doing that separation in a way that makes A1 acceptable to both sides of an argument often

⁴We already know the genetic similarity makes them at least a little more likely to look similar to each other because. For example, siblings are much more likely to have identical DNA test profiles, and two men with identical DNA profiles have the same race[LCMJ10].

⁵See reference. In short, this is the approach of using loosely-related data to set or constrain priors and other parameters, under the assumption that the data is not chosen in a biased way.

requires the description of elaborate (but sufficiently precise) thought experiments that will never be carried out, so I should reiterate that this is not a strategy for directly resolving disputes.

So what is gained? We can now move forward with our disagreement on c_2/c_1 , either by coming to agreement on an A1-type assumption, or by failing to do so. *If we fail to agree on A1, then the quantitative part of the probability assumption was likely obscuring the fundamental source of disagreement more than anything it was contributing. If we agree on A1, then we are left with two (inconsistent) versions of A2-type assumptions, which are as uncertain as the original probability assumption, but are more objective and precise, and thus easier for other people to judge for themselves.*

3.3 Theory comparison

The comparison of theories (or models, explanations of evidence) is a broad category of defeasible reasoning in the physical and social sciences, as well as in criminal law. The Leighton Hay argument (Chapter 6) and the smoking-causes-cancer argument (Chapter 7) are in this category. Both have the following form:

- (a) Specification of one's desired consequence of a comparison being sufficiently-favourable to one's preferred model. For example, that some scientific theory should be abandoned, that some public policy should be put into place, or that some person accused of a crime should be declared guilty.
- (b) Formal definition of when one theory is better than another (or "much better" as in the smoking-causes-cancer argument; the strength needed depends on how much force part (a) requires of the comparison), together with a proposition that part (d) suffices to justify part (a).
- (c) Formalization of the two or more competing theories.
- (d) Deductive proof that one theory is better (or "much better") than the others, according to the definition.

(c) is defeasible in that the formalization of a theory might be unfaithful to the intent of the proponents of the theory, a kind of straw man argument. (b) may be unfair (bias), untrustworthy (variance/bias), or otherwise inappropriate. One type of potential flaw, which is warned about especially often in discussion of the limitations of Bayesian reasoning, is that (i) the definition is a reasonable one based on how well the theories explain or predict the available evidence, but not all evidence relevant to the decision (a) is included. In contrast, the concern that has received the most technical attention

in statistics (both the Bayesian and frequentist schools) is whether (ii) the definition is reasonable *for the available evidence*. Finally, even when all the significantly-relevant evidence is considered and the comparison relation is reasonable, it may be that (iii) the comparison is too weak to justify (a). A critic could try to show any of those three types of flaws in the smoking-causes-cancer or Leighton Hay examples, though I believe types (i) and (iii) would be the most fruitful for them. See Sections 7.1 and 6.3 for some specific criticisms of the smoking-causes-cancer and Leighton Hay arguments, respectively.

3.4 Costs/benefits analysis

There is a common belief in the social sciences that mathematical/logical methods necessarily oversimplify social issues, and because of this they are inappropriate for reasoning about such issues. The first part is true, although it is equally true of natural language argumentation. My main concern here is to argue that the second part is not supported by the first, provided results are reported in a disciplined way.

The Sue Rodriguez argument (Chapter 4) and the physician-assisted suicide argument (Chapter 8) can both be construed as costs/benefits analyses. In both examples, only a subset of the apriori-relevant factors/concerns are considered.

In the Sue Rodriguez argument this is very explicit, since all the goal sentence says is that a certain set of 4 concerns do not justify ruling against her. There remains the possibility that I excluded some concern that is highly relevant, in fact so relevant that with its inclusion there is a strong argument to be made for the negation of the goal sentence *modified to include the 5th concern*. It is tempting to consider my original 4-concern argument invalidated by the hypothetical new 5-concern argument, however that is not technically correct if one treats the arguments in the same way as mathematics arguments; it is not necessarily inconsistent/unreasonable to accept both arguments. Here is an analogy from math: We are interested in whether a majority of elements of a finite set S have a property P . On the way, we prove the answer is no if we replace S with a certain subset $S_1 \subset S$, then later we prove the answer is yes if we replace S with a certain S_2 such that $S_1 \subset S_2 \subset S$, and then no again for a certain S_3 such that $S_1 \subset S_2 \subset S_3 \subset S$. There is no inconsistency, and moreover we are slowly getting closer to the truth about S .

The assisted suicide argument is structured a little differently at its top level, in that its goal sentence is not relative to the simplification (the simplification being that only a subset of the apriori-relevant factors are considered). The argument's Assumption 1 that

Main Lemma⁶ \Rightarrow \langle should pass \rangle ⁷ implicitly includes the simplifying assumption that it is only necessary to consider the direct effects of the assisted suicide system on individual people, and not, for example, on the culture of Canadian society or on groups of people. If a critic goes on to argue the negation of the goal sentence, by including a strictly larger set of factors, but without assuming anything in conflict with the assumptions of my argument about individual people, then they would indeed need to reject Assumption 1 in order to be consistent/reasonable.

It would be very easy to modify the assisted suicide argument so that its goal sentence has the same relative form as the Sue Rodriguez goal sentence⁸, and vice-versa. Moreover, in any case, the goal sentence does not even give the meaning of the proof, which is properly given by $\bigwedge_{A \in \text{Axioms}} A \Rightarrow \text{goal sentence}$. Nonetheless, I think the difference in the two forms of goal sentences is important, if only because of the tendency for proofs to be reported in terms of their main conclusion, with the axioms left tacit. That practice needs to be actively discouraged to answer the oversimplification concern. Fortunately, all that requires is familiarity with FOL and reading the proof document itself. The same cannot be said for defeasible logics, or logics that offer very simple syntax to express complex semantics (see Section 2.3).

3.5 Counterfactual reasoning, hypothetical scenarios

There is a great deal of literature giving general mathematical theories/systems (sometimes called “logics”) for modeling counterfactual reasoning, i.e. reasoning about what would have happened had something been different. Despite the abstract and sometimes-unreal nature of such hypothetical scenarios, we can still reason together deductively, as we are often able to describe such scenarios in such a way that our individual understandings are similar enough that the differences have no significant effect on the argument.

The two arguments about the Berkeley gender bias case (Chapter 5) are examples of deductively-formalized counterfactual reasoning. They have (or can easily be put into) the form:

- (i) Formalize constraints that are (supposedly) sound with respect to an informal theory/explanation/model that one wishes to attack.
- (ii) Formalize constraints that one believes should be required to hold regardless of the theory.

⁶You do not need to lookup what this is for the purpose of this discussion

⁷Which is a 0-ary predicate symbol that means legislation should be passed that introduces an assisted suicide system which is consistent with the constraints given in the argument.

⁸Basically just delete Assumption 1 and make the Main Lemma be the goal sentence

(iii) Deductively show that the constraints together are inconsistent.

It is defeasible in that one's type (i) constraints may misrepresent the informal theory, and in that one's type (ii) constraints may be rejected by opponents. For example, both of the Berkeley arguments make the simplifying (and technically probably wrong) assumption that in the particular pools of applicants to each department, the males and females are equally qualified. That is a type (ii) constraint. The remainder of this section will examine those two arguments in more detail in relation to the itemized form above.

In the first argument, I require a definition of gender prejudice that is weaker than financial forces, in the sense that if all means of discrimination had been removed, then there would not have been a large change in the *total* number of applicants accepted to each department. That is a type (ii) constraint. There is a type (i) constraint that the observed gender biases were caused by prejudice, so that when all means of discrimination are removed, each department should accept men and women at approximately the same rate. Finally, there is the similar type (i) constraint that if all means of discrimination had been removed, it should be possible (under the other constraints) that the overall⁹ acceptance rate for women relative to men improves. Part (iii) consists of proving the negation of that constraint, i.e. that the overall acceptance rate for women gets worse for any admissions round that satisfies the other constraints.

The second argument can alternatively be construed as a theory comparison argument, but I'll describe it in the above form. The main type (i) constraint is effectively that the test used to infer gender discrimination –whether there is a significant bias in favor of men in the overall acceptance rate– should not be sensitive to whether or not the genders apply to different departments at different rates. The main type (ii) assumption is that we can assess that sensitivity by considering an arbitrary pool of applicants of the same size in which men and women apply to each department at close to the same rate, and an arbitrary admissions round in which the gender-specific acceptance rates are close to the same as what they actually were (so if there was gender discrimination in reality, then there should still be gender discrimination in the hypothetical case). In more detail, the main type (i) constraint says that the numeric constraints on the hypothetical admission round should not be enough to guarantee that the test's answer changes from “gender discrimination” to “no gender discrimination”; part (iii) consists of proving the negation of that constraint.

⁹Meaning across all departments.

3.6 Multiplicity of reasons

The most defeasible of defeasible argument types has this form: some n reasons R_1, \dots, R_n are given for a proposition G . It plays a central role in the Carneades system (see Section 9.1.1), for example. Except for the qualitatively-different $n = 1$ case¹⁰ it does not appear in any of the examples in this thesis, and nor should it; the opinion I advocate in this thesis is that, for the problems in the intended problem domain (Section 2.1), the argument form is inappropriate for anything but fast and speculative reasoning.

It is tempting to try to deductively formalize such an argument as

Simplifying Assumption: $R_1 \wedge \dots \wedge R_n \Rightarrow G$ ¹¹

Axiom 1: R_1

...

Axiom n : R_n

However, that formalization is not faithful to the intended meaning of the defeasible argument. It is too fragile. In the framework of deduction, an effective criticism against any one of those $n + 1$ axioms is as good as an effective criticism against them all. In the framework of defeasible reasoning, in contrast, a criticism that effectively argues against only one of the R_i is regarded as only *weakening* the criticized argument. For fast and speculative reasoning, that is a good thing. But when there is adequate time available, the problem is serious, and it is hard to make progress, that principle of defeasible argumentation puts too great a burden on the critic. The burden should be on the argument's author to formalize the sense in which the acceptability of each R_i contributes to the acceptability of G .¹²

Nonetheless, the pattern from the previous paragraph *can be useful* when the axioms are properly interpreted deductively. For example, suppose that the Crown prosecutors of the Canadian government publish a high-level interpreted formal proof that some person is guilty of a murder. A good such argument will employ Bayesian reasoning in some places, but at the top level it could have the above propositional structure. Let's say the goal sentence is a 0-ary predicate symbol G whose language interpretation guide says something along the lines of "the suspect is guilty". The prosecution has evidence linking the murder weapon to the suspect, eye-witness evidence identifying the suspect at the

¹⁰i.e. the general use of conditionals, discussed on page 3.

¹¹Recall from the beginning of this chapter that implication is interpreted classically, so that to accept this simplifying assumption (Section 2.2) means nothing more or less than that you are willing to exclude from your set of personal \mathcal{L} -interpretations any interpretations that satisfy all the R_i but falsify A

¹²Of course, a defeasible logic may provide some sophisticated schemas for formalizing that kind of relationship, but those can just as well be made into reusable first-order theories.

scene of the crime, and DNA evidence of the suspect’s blood collected at the scene of the crime. We’ll make R_1, R_2, R_3 be 0-ary predicate symbols. The prosecution’s language interpretation guide entries for them are:

- R_1 : The murder weapon belonged to the suspect.
- R_2 : The eye-witness correctly saw the suspect fleeing the scene of the crime.
- R_3 : Blood belonging to the suspect was found 10 feet from the victim.

Each R_i is a lemma proved from other assumptions, and they are connected to the goal sentence by the simplifying assumption $R_1 \wedge R_2 \wedge R_3 \Rightarrow G$. This is interesting, surprisingly, when we consider what it means for the prosecution to put forward such a simplifying assumption, and for the defense to accept it. From the prosecution’s perspective, it is useful since it simplifies their task to giving arguments for each of the R_i independently, but it is also risky since it introduces fragility to their argument – the defense only needs to argue against the weakest of the R_i . It is worthwhile for the prosecution if they strongly believe in R_1, R_2, R_3 , and have no other strong inculpatory evidence. From the defense’s perspective, the simplifying assumption is useful since it allows them to focus on refuting the weakest of the R_i , but it is also a concession, since it *is* possible that R_1, R_2, R_3 are true and G is false¹³. It is worthwhile for the defense to accept the simplifying assumption if they think they can give a strong argument that at least one of the R_i is false, and they have no strong exculpatory evidence that isn’t related to the R_i .

¹³For example, it may be that the actual murderer was an associate of the suspect who had access to the suspect’s gun, both were at the scene of the crime (but the murderer was not seen by the witness), and the suspect was injured by the murderer while trying to defend the victim (hence the suspect’s blood).

Chapter 4

Example: Sue Rodriguez's supreme court case

This argument is meant to be read in a browser, and can be found at:

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

I include an inferior static version here just in case you have a printed copy and you strongly prefer to read on paper.

Introduction

This is an argument for granting the right to assisted suicide to a particular individual, as opposed to [an argument for an assisted suicide policy](#), as found in several countries in Europe and a couple American states, and which would provide access to assisted suicide to any Canadian who meets certain requirements. I will adopt a narrative where the party criticizing this proof is the supreme court justices who voted to deny Sue Rodriguez's petition [see [decision](#)]. Exactly the same argument works for the more-recent case of Gloria Taylor, who won her case for access to assisted suicide at the [British Columbia Supreme Court in 2012](#), and lost at the B.C. Court of Appeal in 2013). In Sue Rodriguez'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 (this is [Assumption 7](#)). Thus, the argument comes down to whether allowing Sue Rodriguez ([S.R.](#)) access to assisted suicide would have a negative effect of some sort (against other people - see [Assumption 5](#); or abstract principles - see [Assumption 4](#) and [Assumption 6](#)) that rivals the negative effect of denying her access. The main goals of this argument are:

1. To clarify the qualitative cost to [S.R.](#) of denying her access to legal 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.

One major difference between this formal argument and informal, natural language arguments about assisted suicide cases (those that I've encountered) is the careful distinguishing between

- actions that individuals can do
- states of affairs that individuals want to achieve, which are achievable by their taking certain actions
- states of affairs that we try to prevent by criminalizing certain actions

The primitive symbols of the language of this proof only speak directly about criminalizing actions (sort [Actions](#)). The language speaks indirectly about criminalizing states of affairs (sort [Propositions](#)) via the defined predicate [Justifies criminalizing satisfaction of](#); we can say that the law criminalizes a state of affairs if it criminalizes every action that can achieve that state of affairs. As an example, in many jurisdictions the law indirectly criminalizes any state of affairs ψ in which a young person is high on crack cocaine (\oplus). Suppose I want to justify that law, but without taking a moral stance on whether it is fundamentally wrong to use crack cocaine. Instead, I'll justify it in terms of the desired state of affairs δ that no young person is at risk of becoming *addicted* to crack cocaine. Then I must do two things:

1. Argue that the only actions that can achieve ψ would falsify δ .
2. Argue, or assert, the subjective position that the satisfaction of δ is *more important than* the satisfaction of ψ .

It is important to make these distinctions for this ethical issue for two reasons:

First, because it allows two people to disagree on part of the law while agreeing on a subjective moral position such as (2). For example, it is conceivable that in the future a drug is invented that somehow counteracts the addictive properties of crack cocaine. In such a future, (1) is much easier to reject, and if I reject it then I can argue that δ does not justify criminalizing ψ while still agreeing with (2). In more detail, imagine such an anti-addiction drug is invented, a combination of it and crack cocaine is manufactured, and the combination drug has the property that it is more costly to separate its two component parts than it is to make crack cocaine from scratch. An action a involving the manufacturing and selling of the combination drug is sufficient to attain ψ , but arguably does not risk falsifying δ . In the language of this argument: $a \in \langle \text{actions sufficient for} \rangle(\psi) \wedge \neg \text{Conflicts}(a, \delta)$.

Second, because it prevents one side of the argument from misrepresenting the opinion of the other side. In this example, it prevents supporters of the current law from misrepresenting the opinion of opponents as (being close to) *fundamentally favouring the falsifying of δ* (likely leading to addictions), and it prevents opponents of the current law from misrepresenting the opinion of supporters as (being close to) *fundamentally favouring the impermissibility of ψ* (that being high on crack cocaine is fundamentally wrong). Such misrepresentation happens often in informal argumentation, even sometimes unintentionally!

For the argument below about S.R.'s case, «SueR request» has the role of ψ . Note that its description does not explicitly mention assisted suicide. The conjunction of the elements of sc-concerns has the role of δ . As in the previous paragraph, this prevents misrepresentation of the opinions of the two sides of the issue: The justices who voted against S.R. were *not* arguing that assisting the suicide of another person is fundamentally wrong/impermissible, and the justices who voted in favour of S.R. were *not* arguing that all people have a fundamental right to choose when they will die. However, since I *am* arguing that the court ruled incorrectly, I must connect «SueR request» to the issue of assisted suicide. That is the purpose of \checkmark Lemma 1; it says that the only actions S.R. can take that achieve «SueR request» are ones that involve a physician giving her access to lethal drugs (defined by assisted suicide actions for SueR).

Let's next focus on the **goal sentence** \checkmark Goal of this argument (the final sentence derived from the axioms and lemmas):

$\neg \langle \text{Justifies criminalizing satisfaction of} \rangle(\text{sc-concerns}, \langle \text{SueR request} \rangle)$

sc-concerns is a set of four propositions that were raised by the justices in the majority opinion as desirably-satisfied propositions that might be falsified if they rule in favour of S.R. (\oplus). The goal sentence says that those concerns are not enough to justify criminalizing the satisfaction of «SueR request» (by criminalizing every action that can achieve «SueR request»).

We need a principle of law/morality that connects $\neg \langle \text{Justifies criminalizing}$

satisfaction of $\langle \text{sc-concerns}, \langle \text{SueR request} \rangle \rangle$ and the rest of the proof. This is the predicate $\langle \text{Reach of law limit} \rangle$ (\leftarrow hover cursor to see definition, which will be discussed shortly). I've made it a predicate instead of an axiom to avoid having to posit that the principle holds generally. Instead it is only assumed for one instance, by [Assumption 1](#). That said, I do accept the principle generally with some additional qualification. See the end of the criticism section at the bottom of this page for the improved version $\langle \text{Reach of law limit 2} \rangle$ with additional qualification.

You might think that the defining formula for $\langle \text{Reach of law limit} \rangle$ is surprisingly complicated. There is a simpler, but stronger principle (see definition of $\langle \text{Reach of law limit} \rangle$ for discussion of its weakness), that also suffices to derive $\checkmark \text{Goal}$. It says a set of concerns Δ does not justify criminalizing an *action* a if that action can accomplish a proposition ψ such that for each of the concerns δ , either a does not conflict with δ , or δ is not more important than ψ . I'll now explain why I am reluctant to use this simpler principle. The problem is that ψ is not adequately constrained by a . Consider the contrapositive: If the concerns Δ justify criminalizing an action a , then for *every* proposition ψ that a accomplishes, there must be a concern $\delta \in \Delta$ such that a conflicts with δ and δ is more important than ψ . That consequent is too strong a requirement in some cases! Suppose we want to justify the criminalizing of an action that accomplishes something very good while unnecessarily accomplishing something bad. I'll use an example based on one given by Paul McNamara in a slightly different deontic logic context. We want to argue that it should be criminal to perform the action a of intentionally and unnecessarily breaking a person Timmy's fingers even if it is done while saving Timmy from a fire. The simpler-but-stronger principle that I am reluctant to use says that if we believe that $(*) :=$ "Timmy is saved and his fingers are broken" is more important than "Timmy's fingers are not broken", then we *cannot* consistently justify illegalizing a using just Timmy's desire δ to not have broken fingers. In contrast, with $\langle \text{Reach of law limit} \rangle$ we may consistently criminalize a and believe $(*)$, under the reasonable assumption that it is possible to save Timmy without breaking his fingers.

Argument

Variables $\delta, \psi, \psi_1, \psi_2, \psi_3, \psi_4$ are reserved for sort [Propositions](#).

Variables Δ, Δ' are reserved for sort [Set\(Propositions\)](#).

Variable a is reserved for sort [Actions](#).

Variables X, Y, Z, Y_1, Y_2, Y_3 are reserved for sort [Set\(Actions\)](#).

Variable p is reserved for sort [People](#).

Sort op [Set](#) - Powerset of the given sort.

Sort [Actions](#) \boxminus

Potential concrete actions of individuals. A set each element of which is a concrete/unrepeatable, potential action. By "concrete/unrepeatable", I mean that each action has a definite location (resp. interval of time) where (resp. when) it

would hypothetically occur (\boxplus). Also, each element of this set can be associated with a unique person who performs the action.

Sort **Propositions** \boxminus

Propositions about the real world; things that will turn out true or false. Each will be satisfied or not in every model, but it is nonetheless important that they are distinct from 0-ary predicate symbols, as we will have functions with domain **Propositions**.

Sort **People** \boxminus

The set of residents of Canada who are alive sometime during 1993 (the year of **S.R.**'s supreme court hearing) or later.

Show standard symbols and axioms

Performed : **Actions** $\rightarrow \mathbb{B}$ - The potential actions that are actually performed.

Satisfied : **Propositions** $\rightarrow \mathbb{B}$ - The propositions that turn out true.

S.R. : **People** - Sue Rodriguez

«S.R. facts» : **Propositions** \boxminus

The conjunction of the following list of facts (satisfied propositions) about Sue Rodriguez:

1. **S.R.** has ALS, a usually fatal disease, and multiple doctors have given their opinion that her life expectancy is short.
2. In the late stages of the disease, **S.R.**'s movement will be greatly restricted. If she wishes to live until then (and she does), she will not be able to take her own life without assistance.
3. There is no dispute about whether assisted suicide is *truly* what **S.R.** wants, as evidenced, for example, by the testimony of her friends and family, lack of contradicting testimony from anyone, and her involvement in the Death with Dignity movement.
4. And many more. If some of the assumptions below are expanded into proved lemmas, more facts will be added to this list, and at some point it may be prudent to break this constant up into a number of constants, or a constant of type List(**Propositions**), so that the individual facts can be referred to and criticized more formally.

«SueR request» : **Propositions** \boxminus

Satisfied iff Sue Rodriguez becomes confident that she will be able to take some action (in **Actions**) such that each of the following hold:

1. After completion of the action, she has no further severe pain or indignity caused by her illness.
2. The action, during its execution, does not cause pain or unusual physical or psychological discomfort.
3. Her doing the action does not put any friend or loved one at risk of being convicted of a criminal offence.
4. She does not need to leave Canada permanently in order to do the action.
5. If the action causes permanent loss of consciousness, then it does not occur

until a time when she believes she is no longer able to enjoy life (which she expected would be after losing most of her mobility).

«assisted suicide actions for SueR» : $\text{Set}(\text{Actions}) \sqsubseteq$

The set of actions in which Sue Rodriguez, while in Canada, legally obtains, for the purpose of ending her life, a lethal dose of barbiturates, morphine, or any other drug that is reliably painless, and reliably induces sleeping before it induces loss of consciousness and then death.

a_{SR} : $\text{Actions} \sqsubseteq$

Any element of «assisted suicide actions for SueR». It may be necessary to make it more specific if Assumption 7 is criticized.

«actions sufficient for» : $\text{Propositions} \rightarrow \text{Set}(\text{Actions}) \sqsubseteq$

The set of *all* actions that are feasible and can be expected to result in the given proposition being satisfied.

«More important than» : $\text{Propositions} \times \text{Propositions} \rightarrow \mathbb{B} \sqsubseteq$

A vague (but sufficiently precisifiable) and highly subjective partial order. Example that everyone in the intended audience of this argument should agree on: specific propositions corresponding to instances of "the right to not be murdered" are typically more important than specific propositions corresponding to instances of "the right to free speech".

Conflicts : $\text{Actions} \times \text{Propositions} \rightarrow \mathbb{B} \sqsubseteq$

If the relation holds for (a, ψ) then a conflicts with ψ in the sense that if a is performed then (causally, directly or indirectly) ψ cannot be satisfied.

not : $\text{Propositions} \rightarrow \text{Propositions} \sqsubseteq$

The proposition that is satisfied iff the given proposition is not satisfied.

«regrettable legal assisted suicide» : $\text{People} \rightarrow \mathbb{B} \sqsubseteq$

«regrettable legal assisted suicide»(p) is the proposition that is satisfied iff:

1. p dies by a legally-sanctioned use of assisted suicide.
2. There is some information about p , which was unknown at the time when their application for assisted suicide was approved, that, if it had been known, would have caused a significant proportion (say, 5%) of people who would have supported p 's application to resolutely change their mind. Here "resolutely" means that no further information about p would again change the minds of those 5% of people. We more simply (but slightly more vaguely) say that at least 5% of p 's supporters, if given "perfect" information about p , would change their minds,

«avoid judicial overreach» : $\text{Propositions} \sqsubseteq$

The ruling of the supreme court justices on Sue Rodriguez's case does not constitute "judicial overreach".

«avoid legal precedent causing slippery slope» : $\text{Propositions} \sqsubseteq$

Satisfied if «SueR request» is not satisfied, or if «SueR request» is satisfied and a certain kind of "slippery slope" is blocked; in particular, permitting S.R.'s assisted suicide request does not "lead to" the Supreme Court or a lower court permitting a regrettable (see «regrettable legal assisted suicide») instance of assisted suicide.

Defn~ «no regrettable legal assisted suicide for S.R.» : Propositions - The proposition that is satisfied iff «regrettable legal assisted suicide»(S.R.) is not satisfied. \square

Satisfied(«no regrettable legal assisted suicide for S.R.») $\Leftrightarrow \neg$ «regrettable legal assisted suicide»(S.R.)

«consistency with maj opinion» : Propositions \square

The proposition that is satisfied iff the decision made by the Supreme Court is "in agreement with the majority opinion" of Canadian citizens on whether S.R. should be granted an exception to the criminal code.

Defn sc-concerns : Set(Propositions) - "sc" for Supreme Court. Some of the concerns of opponents of assisted suicide, formulated as propositions that they want to be true. Specifically, they are the concerns mentioned in the majority opinion for the actual Supreme Court decision. \square

sc-concerns = {«avoid judicial overreach», «avoid legal precedent causing slippery slope», «no regrettable legal assisted suicide for S.R.», «consistency with maj opinion»}

«Justifies criminalizing» : Set(Propositions) \times Set(Actions) $\rightarrow \mathbb{B}$ \square

The desirability of satisfying the given propositions justifies criminalizing the given actions.

Defn «Justifies criminalizing satisfaction of»(Δ, ψ) : Set(Propositions) \times Propositions $\rightarrow \mathbb{B}$ - The desirability of satisfying the given propositions Δ justifies criminalizing the satisfaction of ψ . \square

$\forall \Delta, \psi$. «Justifies criminalizing satisfaction of»(Δ, ψ) \Leftrightarrow «Justifies criminalizing»(Δ , «actions sufficient for»(ψ))

Defn «Reach of law limit» : Set(Propositions) \times Propositions $\rightarrow \mathbb{B}$ - The defining formula of this predicate is a general, but weak, principle of liberalism, which we will use one instance of (Assumption 1). It only requires justification for laws that criminalize all possible ways of accomplishing a proposition ψ , saying nothing about laws that criminalize, *without* justification, *some but not all* actions than can accomplish ψ .

It says that the set of (ostensibly desired) propositions Δ does *not* justify criminalizing the satisfaction of the proposition ψ (by criminalizing all the actions that can accomplish ψ) if there is an action that can accomplish ψ such that, for each of the desired propositions $\delta \in \Delta$, either δ is not more important than ψ , or the action does not conflict with δ .

Contrapositive: Suppose that a set of (ostensibly desired) propositions Δ justifies criminalizing the set of all actions that can accomplish another proposition ψ . Intuitively, this means Δ justifies criminalizing the satisfaction of ψ . Then, it must be that for each of those actions a that can achieve ψ , there is a proposition $\delta \in \Delta$ that both conflicts with a and is more important than ψ . \square

$\forall \Delta$. $\forall \psi$. «Reach of law limit»(Δ, ψ) $\Leftrightarrow ((\exists a \in$ «actions sufficient for»(ψ)). $\forall \delta \in \Delta$. \neg «More important than»(δ, ψ) $\vee \neg$ «Conflicts»(a, δ)) $\Rightarrow \neg$ «Justifies criminalizing satisfaction of»(Δ, ψ))

□ ✓Goal: The specific set of concerns *sc-concerns* does not justify criminalizing the satisfaction of «SueR request». □

¬⟨Justifies criminalizing satisfaction of⟩(*sc-concerns*, «SueR request»)

The theorem is a logical consequence of the following axioms; the code that generates this HTML file also generates first-order validity problems, which were solved by CVC4 and Vampire via [System on TPTP](#). Each axiom can be disputed, and some, with more work, can be made into lemmas, proved from more-basic assumptions and simplifying assumptions. Each axiom is informally labeled an Assumption or Assertion. The Assertions are intended to be uncontroversial.

Assumption 1: Suppose there is some action *a* that *S.R.* can take to achieve «SueR request» such that, for any concern $\psi \in \textit{sc-concerns}$ that is *not* strictly less important than «SueR request», the action *a* does not actually conflict with ψ . Then *sc-concerns* does not justify criminalizing the set of all actions that can achieve «SueR request». □

⟨Reach of law limit⟩(*sc-concerns*, «SueR request»)

□ ✓Lemma 1: There are no actions that can achieve *S.R.*'s request other than the ones described above in ⟨assisted suicide actions for SueR⟩ (all of which involve her use of assisted suicide). □

⟨assisted suicide actions for SueR⟩ = ⟨actions sufficient for⟩(«SueR request»)

Argument sketch: I claim that there are three broad categories of actions that might plausibly be able to achieve the satisfaction of «SueR request»: (1) treatment, (2) suicide, or (3) palliative sedation (aka terminal sedation). In Rodriguez's and Taylor's cases, there are no sufficient treatments for ALS and there is no hope for the discovery and availability of a new one before their death, so (1) is out. To dismiss (3) one must do some reading, e.g. [Palliative Sedation: It's Not a Panacea](#); in short, the ideal of terminal sedation, in which a dying patient's life is not shortened, but all their suffering is medicated away, is far from achieved in practice. If it *was* ideal, we would include in ⟨actions sufficient for⟩(«SueR request») actions by which a patient is guaranteed access to terminal sedation (there is currently no general way of getting such a guarantee in Canada; one has to just get lucky to end up with a doctor who is willing to do it).

$A_t, A_s, A_{ps} : \text{Set}(\text{Actions})$ □

Actions that could plausibly achieve, respectively, a treatment/cure, suicide (including assisted suicide), or palliative sedation, for *S.R.*

□ **Assertion 1:** Every action that can achieve *S.R.*'s request is one involving a treatment/cure of her condition, some form of suicide, or some form of palliative sedation. □

⟨actions sufficient for⟩(«SueR request») $\subseteq A_t \cup A_s \cup A_{ps}$

Argument sketch: I claim that there are three broad categories of actions that might plausibly be able to achieve the satisfaction of «SueR request»:

(1) treatments/cures, (2) suicide, or (3) palliative sedation (aka terminal sedation). I don't anticipate that this would be disputed.

□ **Assertion 2:** Treatment/cure-seeking actions cannot achieve S.R.'s request. □

$$\langle \text{actions sufficient for} \rangle(\langle \text{SueR request} \rangle) \cap A_t = \emptyset$$

Argument sketch: While ALS could be effectively treated or cured some day, «SueR request» would require the discovery and minimal testing of such a treatment within a year or two, and there is negligible hope within the medical community for that.

□ **Assumption 2:** Palliative sedation cannot achieve S.R.'s request. □

$$\langle \text{actions sufficient for} \rangle(\langle \text{SueR request} \rangle) \cap A_{ps} = \emptyset$$

Argument sketch: See [Palliative Sedation: It's Not a Panacea](#); in short, the ideal of terminal sedation, in which a dying patient's life is not shortened, but all their suffering is medicated away, is far from achieved in practice. If it *was* ideal, we would need to include, in $\langle \text{actions sufficient for} \rangle(\langle \text{SueR request} \rangle)$, actions through which a patient becomes guaranteed access to terminal sedation. However, there is currently no general way of getting such a guarantee in Canada; one has to just get lucky to end up with a doctor who is willing to do it.

□ **Assumption 3:** Among the possible ways that S.R. could end her own life, only those in $\langle \text{assisted suicide actions for SueR} \rangle$ satisfy the criteria of «SueR request». □

$$\langle \text{actions sufficient for} \rangle(\langle \text{SueR request} \rangle) \cap A_s = \langle \text{assisted suicide actions for SueR} \rangle$$

Argument sketch: Proving this involves a morose consideration of all the known methods of suicide, observing that each of them, besides the use of legally-prescribed sedatives, violates at least one of the conditions of «SueR request».

Assertion 3: The uncontroversial assertion that the actions described in $\langle \text{assisted suicide actions for SueR} \rangle$ would suffice to meet S.R.'s desired condition «SueR request». □

$$\langle \text{assisted suicide actions for SueR} \rangle \subseteq \langle \text{actions sufficient for} \rangle(\langle \text{SueR request} \rangle)$$

Assertion 4: The specific action a_{SR} is in $\langle \text{assisted suicide actions for SueR} \rangle$ (informally "by definition"). □

$$a_{SR} \in \langle \text{assisted suicide actions for SueR} \rangle$$

□ **Assumption 4:** The judges' necessary involvement in making it legal for S.R. to obtain lethal prescription drugs need not constitute judicial overreach. □

□ $\neg \text{Conflicts}(a_{SR}, \langle \text{avoid judicial overreach} \rangle)$

Argument sketch: Claim that a suspended annulment, with a period of at least one year, plus a special waiver for one person, is *never* judicial overreach. In a suspended annulment, a law is declared unconstitutional, but it is allowed to remain in effect for a period of time, to give the legislative branch the opportunity to replace it with a new, constitutional law. Reading: [Myth of](#)

Judicial Overreach

☐ **Assumption 5:** The judges' necessary involvement in making it legal for **S.R.** to obtain lethal prescription drugs need not create a legal precedent that leads to a "slippery slope". ☐

¬**Conflicts**(a_{SR} , «avoid legal precedent causing slippery slope»)

Argument sketch: Though the justices may not be able to artificially specify that their ruling in favour of **S.R.** should not be used as precedent, they can certainly restrict the extent of the precedent, by specifying only that the law is unconstitutional for any citizen satisfying «**S.R. facts**». A citizen petitioning a lower court for access to assisted suicide, who does not satisfy all of «**S.R. facts**», would be neither helped nor hindered by the ruling in favour of **S.R.** Given that consideration, we can use the same argument that we use for **Assumption 7** to justify this assumption (since in that argument «**S.R. facts**» are the only facts we use about **S.R.**).

☐ **Assumption 6:** **S.R.**'s claim for «**SueR request**» trumps her opponents' claim for Supreme Court decisions to be consistent with the majority opinion (among Canadian citizens). ☐

¬**More important than**(«consistency with maj opinion», «**SueR request**»)

Argument sketch: The goal of protection against tyranny of the majority is precisely what makes the constitution special, compared to other laws. My opinion is that «consistency with maj opinion» should be given very small (if not zero) weight when assessing whether a part of the law should be repealed on constitutional grounds. To replace this assumption with a high-level proof... ☐

☐ **Assumption 7:** The particular assisted suicide acton a_{SR} that we chose does not conflict with the desire to avoid **S.R.** being the victim of a «regrettable legal assisted suicide». ☐

¬**Conflicts**(a_{SR} , «no regrettable legal assisted suicide for **S.R.**»)

Argument sketch: This is easily argued by reference to «**S.R. facts**».

Hide sample criticism

Criticizing the argument

The justices that ruled against Sue Rodriguez (hereafter called "the majority", as in the decision itself) and wrote the majority opinion in [the Rodriguez ruling](#) implicitly criticized [Assumption 4](#), [Assumption 5](#), and [Assumption 6](#), but not [Assumption 7](#). Their arguments did not significantly touch on the details considered in [✓Lemma 1](#), which is not surprising as they did not attempt to reason with non-trivial precision about the costs and benefits for Sue Rodriguez.

Take [Assumption 6](#) for example. The authors of the majority opinion write: "*the issue before the Court was whether a criminal prohibition on assisting suicide in situations*

where a person is terminally ill and mentally competent but unable to commit suicide by him or herself, is contrary to the principles of fundamental justice. What are principles of fundamental justice? Mr. Justice Sopinka noted that determining these principles can be an onerous task. Such principles, he pointed out, are those for which there is some consensus among reasonable people as to their importance to our societal concept of justice."

From that statement, we can see that the majority rather flatly disagree with the opinion I express in the prose below the statement of [Assumption 6](#). So how to proceed after that? I could dispute their claim about the consensus of the Canadian public (probably hard, but that situation is slowly improving in favour of assisted suicide), or I could look for other cases that demonstrate the justices informally contradicting their allegiance to the principle expressed in the above quote (probably easy), or I could take this as a fundamental source of subjective disagreement. For the last option, I would move on to their criticisms of the other assumptions, and try to refute them (by demonstrating informal inconsistency), so that we are left with only one source of fundamental disagreement.

Another possible criticism is that [sc-concerns](#) is too small a set. A critic may wish to add an entirely new [Propositions](#) constant to [sc-concerns](#), or perhaps the conjunction of two or more of the current elements of [sc-concerns](#). An addition of the latter type could conceivably have a significant effect if combined with an addition of the first type, since [Assumption 1](#) considers the elements of the first argument to [Reach of law limit](#) separately (an addition of the first type would be necessary to really force me to respond in a challenging way, since currently only one of the 4 primitive concerns are compared to [SueR request](#) with [More important than](#)). In any case, a language modification and extension that makes [Reach of law limit](#) respect conjunctions of elements of [sc-concerns](#) is as follows:

and : $\text{Set}(\text{Propositions}) \rightarrow \text{Propositions}$ - Conjunction of the given set of propositions.

\subseteq : $\text{Set}(\text{Propositions}) \times \text{Set}(\text{Propositions}) \rightarrow \mathbb{B}$ - Subset

\setminus : $\text{Set}(\text{Propositions}) \times \text{Set}(\text{Propositions}) \rightarrow \text{Set}(\text{Propositions})$ - Set difference

Defn [Reach of law limit 2](#) : $\text{Set}(\text{Propositions}) \times \text{Propositions} \rightarrow \mathbb{B}$ - [Reach of law limit 2](#)(Δ, ψ) asserts the following statement, which is an implication. If Δ can be partitioned into two sets Δ' and $\Delta \setminus \Delta'$ such that

- the conjunction of the concerns Δ' is not more important than ψ , and
- there is an action a that can accomplish ψ such that none of the concerns in $\Delta \setminus \Delta'$ actually conflict with a

then Δ does not justify criminalizing all the actions that can accomplish ψ . \square

$\forall \Delta, \psi. \langle \text{Reach of law limit 2} \rangle(\Delta, \psi) \Leftrightarrow ((\exists \Delta' \subseteq \Delta. \neg \langle \text{More important than} \rangle(\text{and}(\Delta'), \psi) \wedge (\exists a \in \langle \text{actions sufficient for} \rangle(\psi). \forall \delta \in \Delta \setminus \Delta'. \neg \text{Conflicts}(a, \delta))) \Rightarrow \neg \langle \text{Justifies criminalizing satisfaction of} \rangle(\Delta, \psi))$

Chapter 5

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.¹ 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.

Department	Male		Female		Total	
	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],² though it makes weaker assumptions (Assumption 8 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

¹The 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.

²The 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.

discrimination” and reasoning from there. More specifically, it first fixes a definition of “gender discrimination”, and then defines (in natural language) a hypothetical admissions protocol that prevents gender discrimination by design. Considering then a hypothetical round-of-admissions scenario that has the same set of applications as in the actual round of admissions, if we assume that the ungendered departmental acceptance rates are not much different in the hypothetical scenario, then it can be shown that the overall bias is actually *worse* for women in the hypothetical scenario. Since the hypothetical scenario has no gender discrimination by design, and is otherwise as similar as possible to the real scenario, we conclude that the observed bias cannot be blamed on gender discrimination.

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

I’ll use g to refer to the *language interpretation guide* for the language \mathcal{L} of this argument.

$\mathcal{L} \setminus \mathcal{L}_{\text{rigid}}$ consists of:

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

$\mathcal{L}_{\text{rigid}}$ consists of:

- A number of mathematical symbols that have their standard meaning: constants $0, 1, 512, 825, \dots$, function symbols $|\cdot|, \cap, \cup, +, -, *, /$, predicate symbols $<, =$.
- The constants $\text{App}, \text{Acc}, \text{App}^m, \text{App}^f, \text{App}^1, \dots, \text{App}^6$. 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 sorts are \mathcal{A} for application sets and \mathbb{Q}^\perp for the rational numbers with an element for “undefined”. See below for g ’s entries for them.

The types of the function/predicate symbols other than the 0-ary predicate symbols (and besides $=$, which is untyped) are as follows. With respect to the definition of

interpreted formal proof from Section 2.2, they are all *assumptions* as opposed to *simplifying assumptions*.

$$\begin{aligned}
 \text{App, Acc, Acc}_{\text{hyp}}, \text{App}^m, \text{App}^f, & : \mathcal{A} \\
 \text{App}^1, \dots, \text{App}^6 & \\
 |\cdot| & : \mathcal{A} \rightarrow \mathbb{Q}^\perp \\
 0, 1, 512, 825, \dots & : \mathbb{Q}^\perp \\
 \cap, \cup & : \mathcal{A} \times \mathcal{A} \rightarrow \mathcal{A} \\
 +, -, * & : \mathbb{Q}^\perp \times \mathbb{Q}^\perp \rightarrow \mathbb{Q}^\perp \\
 / & : \mathbb{Q}^\perp \times \mathbb{Q}^\perp \rightarrow \mathbb{Q}^\perp \\
 < & : \mathbb{Q}^\perp \times \mathbb{Q}^\perp \rightarrow \mathbb{B}^4
 \end{aligned}$$

5.1 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.

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

$g(\langle \text{bias only evidence} \rangle)$ 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.”

$g(\langle \text{gender uncor with ability in each dept} \rangle)$ 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]

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

$g(\mathbb{Q}^\perp) =$ The rational numbers plus one extra object for error/undefined.

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

g also says that

- $0, 1, 512$, etc are the expected numerals.
- $|\cdot|$ is the function that gives the size of each set in \mathcal{A} .
- \cap, \cup are the expected binary functions on \mathcal{A} .
- $+, -, *$ are the expected binary functions on the naturals extended so that they equals \perp when either or both of the arguments are \perp .
- $/$ is division on the rationals extended so that it equals \perp iff one or both of the arguments are \perp or the second argument is 0.
- $<$ is the usual ordering on the rationals extended by making \perp be neither greater than nor less than any number or itself.

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

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

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

$g(\mathbf{Acc}_{\text{hyp}}) =$ 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 elaborate⁵, 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, $\mathbf{Acc}_{\text{hyp}}$ is the set of accepted applications in the hypothetical scenario.

$g(\mathbf{App}^m) = \mathbf{App}^m$ 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.

$g(\mathbf{App}^f) = \mathbf{App}^f$ is a subset of **App** of size 1835 (sum of the second “Applied” column in

⁵It need not be efficient/economical, since we are only introducing the scenario as a reasoning device.

the table), specifically the applications where the applicant is female.

For $d = 1, \dots, 6$:

$g(\text{App}^d) = \text{App}^d$ is the set of applications for admission into department d .

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

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

$$\text{App}^{d,g} := \text{App}^d \cap \text{App}^g$$

$$\text{Acc}^{d,g} := \text{App}^{d,g} \cap \text{Acc}$$

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

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

Assumption 7. 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 $d \in \{1, \dots, 6\}$ and $g \in \{m, f\}$:

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

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

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

Assumption 8. 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 $d \in \{1, \dots, 6\}$:

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

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

The constants .037 and 9 are roughly the most extreme values that make the proof go through. To illustrate the first form, the bounds for the departments with the fewest and greatest number of accepted applicants are:

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

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

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

Assumption 9. If $\langle \text{bias only evidence} \rangle$ and

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

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

Simplifying Assumption 2. $\langle \text{bias only evidence} \rangle$

Claim 1.

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

Proof. It is not hard to formulate this as a linear integer programming problem, where the variables are the sizes of the sets $\text{Acc}_{\text{hyp}}^{d,g}$. Coming up with inequalities that express the previous axioms and the data axioms from Section 5.3 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. \square

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

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

5.2 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 $\text{App}_{\text{hyp}}^d : \mathcal{A}$ for $1 \leq d \leq 6$.

The hypothetical admissions round is similar to the true admissions round (Axioms 10 and 12) except that men and women apply to each department at close to the same rate (Assumption 11) - 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 $g(\text{App}_{\text{hyp}}^d)$ and $g(\text{Acc}_{\text{hyp}})$ to reflect these alternate assumptions.

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

Assumption 10. 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.⁶

For $d \in \{1, \dots, 6\}$ and $g \in \{m, f\}$:

$$|\text{App}_{\text{hyp}}^d| = |\text{App}^d|, \quad |\text{App}_{\text{hyp}}^g| = |\text{App}^g|$$

Assumption 11. In the hypothetical scenario, gendered departmental *application* rates are close to gender-independent. For $d \in \{1, \dots, 6\}$ and $g \in \{m, f\}$:

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

Assumption 12. In the hypothetical scenario, gendered departmental *acceptance* rates are close to the same as in the real scenario.

For $d \in \{1, \dots, 6\}$ and $g \in \{m, f\}$:

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

Claim 3. $\text{accRate}_{\text{hyp}}^f > \text{accRate}_{\text{hyp}}^m$

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

⁶This 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.

axioms from the next section is easy. Then, add the constraint

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

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

Assumption 13. If $\langle \text{bias only evidence} \rangle$ and $\text{accRate}_{\text{hyp}}^f > \text{accRate}_{\text{hyp}}^m$ then $\langle \text{lawsuit should be dismissed} \rangle$

Simplifying Assumption 2 from the previous proof, which just asserts $\langle \text{bias only evidence} \rangle$, is also used here. From it, Assumption 13, and Claim 3, the goal sentence $\langle \text{lawsuit should be dismissed} \rangle$ follows immediately.

5.3 Data Axioms

Assumption 14.

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

$$|\text{App}^{1,m}| = 825, \quad |\text{Acc}^{1,m}| = 512, \quad |\text{App}^{1,f}| = 108, \quad |\text{Acc}^{1,f}| = 89$$

$$|\text{App}^{2,m}| = 560, \quad |\text{Acc}^{2,m}| = 353, \quad |\text{App}^{2,f}| = 25, \quad |\text{Acc}^{2,f}| = 17$$

$$|\text{App}^{3,m}| = 325, \quad |\text{Acc}^{3,m}| = 120, \quad |\text{App}^{3,f}| = 593, \quad |\text{Acc}^{3,f}| = 202$$

$$|\text{App}^{4,m}| = 417, \quad |\text{Acc}^{4,m}| = 138, \quad |\text{App}^{4,f}| = 375, \quad |\text{Acc}^{4,f}| = 131$$

$$|\text{App}^{5,m}| = 191, \quad |\text{Acc}^{5,m}| = 53, \quad |\text{App}^{5,f}| = 393, \quad |\text{Acc}^{5,f}| = 94$$

$$|\text{App}^{6,m}| = 373, \quad |\text{Acc}^{6,m}| = 22, \quad |\text{App}^{6,f}| = 341, \quad |\text{Acc}^{6,f}| = 24$$

That App is the disjoint union of $\text{App}^1, \dots, \text{App}^6$ follows from the previous sentences (under the standard interpretation of numbers and sets).

Chapter 6

Example: Leighton Hay's wrongful 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)¹, than should have been necessary to convict. A good, short summary about the case can be found here:

<http://www.theglobeandmail.com/news/national/defence-prosecution-split-on-need-for-forensic-hair-testing/article1367543/>

The prosecution's case relies strongly on the testimony of one witness, Leisa Maillard, who picked (a 2 year old picture of) Hay out of a photo lineup of 12 black men of similar age, and said she was 80% sure that he was the shooter. There were a number of other witnesses, none of whom identified Hay as one of the killers. Ms. Maillard'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 the following morning. 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

¹Thanks to Joanne McLean and Deryck Ramcharitar for making the case files available to me.

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 6.1: 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 bin₃ and bin₄ are very likely beard hairs, and that the hairs in bin₁ and bin₂ could be either beard or scalp hairs.

remained stuck to the newspaper (Due perhaps to being lighter than the dreads? It was not explained why.). It is the origin of those hair clippings that we are concerned with in this argument; 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. Maillard’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 that we use in this argument. ~~We do not yet know whether the Crown will choose to prosecute Hay again, or if they do, whether they will attempt to again use the hair clippings as evidence against him.~~ On 28 Nov 2014, the Crown dropped its murder charges against Hay, declining to prosecute him again, and he was freed. As usual in these cases, there was no pronouncement of innocence, and Hay and his lawyers will have to fight for monetary compensation.

6.1 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

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 6.2: 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.

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

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

1. There are *many* distinct theories of how the hypothesized beard and scalp trims could have happened. In the argument below, we introduce a family of such theories indexed by the parameters 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 (supplied) reasonable justification (without referring to the measurements, which would be using the data to fit the model that the data is supposed to predict). 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 only superficial changes to the axioms.

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

G, Clipped, Mix, BParams, H, Widths

Some of the most significant assumptions for the argument are as follows:

1. The prior chosen for the suspect's beard hair-width distribution is fair and reasonable.² This is Simplifying Assumption 4. It is probably the most disputable of the assumptions. I give some criticisms of it in Section 6.3.
2. The distribution for the suspect's scalp hair widths, based on the samples taken in 2010, is fair and reasonable (Simplifying Assumption 6). This assumption may be disputable in that it does not assume that Hay's scalp hairs thinned by an average amount for a man of his age and race in the 8 years between the crime and when his scalp hair sample was taken. Of course, it may be that his hairs have not thinned at all. Unfortunately it appears that we cannot know this, as samples were not taken in 2002.
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 3) from a sufficiently large range the ratio

$$\frac{\text{P}(\text{random clipped hair came from beard, given only that it ended up in the newspaper})}{\text{P}(\text{random clipped hair came from the scalp, given only that it ended up in the newspaper})} \quad (6.1)$$

Specifically that range is $[\frac{\alpha_{\min}}{1-\alpha_{\min}}, \frac{\alpha_{\max}}{1-\alpha_{\max}}]$ (but note: no axiom of the argument requires that symbolic form to be meaningful to the reader). The axioms enforce no constraints about α_{\min} and α_{\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

²The 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.

$\frac{P(\text{Widths}=\bar{b}|\mathbb{G})}{P(\text{Widths}=\bar{b}|\bar{\mathbb{G}})}$ to be ≥ 1 , the prosecution needs to make an extreme assumption about α_{\min} and α_{\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 (6.1). 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 22. 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 results in a version of Claim 6 that is *better* for the defense. I don’t do this simply because the extra complexity would reduce the pedagogical value of this example.

6.2 Argument

Because this argument is written in L^AT_EX, I present it more-informally than is required by the definition *interpreted formal proof*. In particular, I do not explicitly name the types of most symbols, and I don’t explain how exactly random variables and the P(proposition | proposition) syntax are formalized.

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 $\mathcal{L}_{\text{rigid}}$, 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:

$$P(t_1, t_2 | t_3) = P(t_1 | t_2, t_3)P(t_2 | t_3)$$

- 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 :

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

bin₁, **bin**₂, **bin**₃, **bin**₄ are constants denoting the four micrometer-intervals from Table 1. Formally, they belong to their own sort, which has exactly 4 elements in every model. We do not actually have micrometer intervals in the ontology of the proof, so we could

just as well use $\{1, 2, 3, 4\}$, but I think that would be confusing later on. **Bins** is the sort $\{\text{bin}_1, \text{bin}_2, \text{bin}_3, \text{bin}_4\}$.

Throughout this writeup, $\vec{b} = b_1, \dots, b_{89}$ is a fixed ordering of the newspaper hair measurements shown in Table 1. Specifically, each b_i is one of the constants $\text{bin}_1, \text{bin}_2, \text{bin}_3$, or bin_4 ; bin_1 appears 10 times, bin_2 20 times, bin_3 40 times, and bin_4 19 times.

\vec{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 $\mathbf{G} = \text{true}$ and $\overline{\mathbf{G}}$ to abbreviate $\mathbf{G} = \text{false}$.

Clipped is a simulation random variable whose value is determined by G . When G is false, **Clipped** is the set of beard hair fragments that fall from the suspect’s face when he does a full beard trim with an electric trimmer³ several days before the murder took place. When G is true, **Clipped** is the set of beard and scalp hair fragments that fall from the suspect’s head when he does a full beard trim and a full scalp trim (the latter after cutting off his two-inch dreds) with the same electric trimmer. This includes any such fragments that were flushed down the sink or toilet, but not including—in the case that the suspect is guilty—hair fragments that were part of his 2-inch “picky dreads.”

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

BParams is the simulation random variable that gives the parameters of the suspect’s beard hair width distribution.

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

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

³The 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.

respectively. Similarly, I write just α as an abbreviation for $\text{Mix} = \alpha$.

\mathbf{B} is the set from which our prior for the suspect's beard hair width distribution is defined. It is the set of triples $\langle p_1, p_2, p_3 \rangle \in [0, 1]^3$ such that $p_1 \leq p_2, p_3, p_4$ 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 .

$P(t_1 \mid t_2)$ is the notation we use for the Bayesian/simulation distribution over the random variables \mathbf{G} , Clipped , Mix , BParams , \mathbf{H} , 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.

NOTATION: 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: $\forall \vec{p} \in \mathbf{B}$ and $\forall \alpha \in [\alpha_{\min}, \alpha_{\max}]$.

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

Definition 4. We are aiming to show that from reasonable assumptions, the following likelihood ratio is less than 1, meaning that the defense's theory explains the newspaper hairs evidence at least as well as the prosecution's theory. The notation *likelihood-ratio* $(\alpha_{\min}, \alpha_{\max})$ is used just to highlight the dependence on the parameters $\alpha_{\min}, \alpha_{\max}$.

$$\textit{likelihood-ratio}(\alpha_{\min}, \alpha_{\max}) := \frac{P(\text{Widths} = \vec{b} \mid \mathbf{G})}{P(\text{Widths} = \vec{b} \mid \overline{\mathbf{G}})}$$

Assumption 15. The values of BParams and Mix are chosen independently of each other and \mathbf{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 $t \in \{\text{true}, \text{false}\}$:

$$d_{(\text{BParams}, \text{Mix})}(\vec{p}, \alpha \mid \mathbf{G} = t) = d_{\text{BParams}}(\vec{p}) \cdot d_{\text{Mix}}(\alpha)$$

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

Simplifying Assumption 3. The prior distribution for the mixture parameter Mix is the uniform distribution over $[\alpha_{\min}, \alpha_{\max}]$.

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

Simplifying Assumption 4. The prior distribution for the parameters of the suspect's beard hair width distribution is the uniform distribution over the set $B \subseteq [0, 1]^3$ defined above.

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

$\text{News}(h) = \text{true}$ iff the hair clipping h ends up in the balled-up newspaper.

$\text{Beard}(h) = \text{true}$ (respectively $\text{Scalp}(h) = \text{true}$) iff hair clipping h came from the suspect's beard (respectively scalp).

Assumption 16. Both prosecution and defense agreed that all the hairs in the newspaper came from the suspect's beard or scalp, and not both.⁴

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

width is the function from 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 5. In the simulation model, the hairs that ended up in the newspaper are chosen independently at random with replacement from some hair-width distributions.

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

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

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

⁴“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.

newspaper hair is a beard hair.

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

Proof. Follows from Assumption 16. □

Assumption 17. In the defense's model (not guilty $\bar{\mathbf{G}}$), all the newspaper hair came from a beard trim, and so the mixture parameter is irrelevant.

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

Assumption 18. 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.

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

Assumption 19. 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.

$$\begin{aligned} \alpha &= P(\text{Beard}(\text{H}) \mid \text{News}(\text{H}), \mathbf{G}, \vec{p}, \text{Mix} = \alpha) \\ 1 - \alpha &= P(\text{Scalp}(\text{H}) \mid \text{News}(\text{H}), \mathbf{G}, \vec{p}, \text{Mix} = \alpha) \end{aligned}$$

Assumption 20. The precise meaning of the simulation random variable BParams . Recall that p_4 abbreviates $1 - p_1 - p_2 - p_3$. For $j \in \{1, 2, 3, 4\}$:

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

Simplifying Assumption 6. 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.

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

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

Assumption 21. If $\frac{P(\text{Widths}=\vec{b}|\mathbf{G})}{P(\text{Widths}=\vec{b}|\overline{\mathbf{G}})} \leq 1$ (i.e. *likelihood-ratio* ≤ 1), then \langle the newspaper hair evidence is neutral or exculpatory \rangle .⁵

Claim 5. If $\alpha_{\min} \leq .849$ then $\frac{P(\text{Widths}=\vec{b}|\mathbf{G})}{P(\text{Widths}=\vec{b}|\overline{\mathbf{G}})} < 1$

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 22, the ratio on the left side being the new parameter), we will obtain a corollary of Claim 5 that is easier to interpret.

We do not know what the ratio of beard to scalp hairs on Hay’s head was on the date of the murder, and it is not hard to see that a higher value of $P(\text{Beard}(\text{H}) \mid \mathbf{G}, \vec{p}, \alpha)$ is favourable for the prosecution.⁶ 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).⁷ Thus we assume:

Assumption 22.

$$\frac{P(\text{Beard}(\text{H}) \mid \mathbf{G}, \vec{p}, \alpha)}{P(\text{Scalp}(\text{H}) \mid \mathbf{G}, \vec{p}, \alpha)} \leq 4/11$$

Claim 6. The hypothesis of Assumption 21 also follows if we assume Assumption 22 and that *the uniform prior over Mix gives positive density to a model where a random clipped beard hair is ≤ 15 times more likely to end up in the newspaper as a random clipped scalp hair*:

⁵The text in brackets is a constant predicate symbol.

⁶Raising the value makes both models worse, but it hurts the prosecution’s model less since the prosecution’s model can accommodate by lowering α_{\min} and α_{\max} .

⁷Trustworthy 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.

If there exists $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ and $\vec{p} \in B$ such that

$$\frac{P(\text{News}(\text{H}) \mid \text{Beard}(\text{H}), \mathbf{G}, \vec{p}, \alpha)}{P(\text{News}(\text{H}) \mid \text{Scalp}(\text{H}), \mathbf{G}, \vec{p}, \alpha)} \leq 15$$

then

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

Proof. Let α, \vec{p} be as in the hypothesis.

From basic rules about conditional probabilities:

$$\frac{\alpha}{1 - \alpha} = \frac{P(\text{Beard}(\text{H}) \mid \text{News}(\text{H}), \mathbf{G}, \vec{p}, \alpha)}{P(\text{Scalp}(\text{H}) \mid \text{News}(\text{H}), \mathbf{G}, \vec{p}, \alpha)} = \frac{P(\text{News}(\text{H}) \mid \text{Beard}(\text{H}), \mathbf{G}, \vec{p}, \alpha)}{P(\text{News}(\text{H}) \mid \text{Scalp}(\text{H}), \mathbf{G}, \vec{p}, \alpha)} \frac{P(\text{Beard}(\text{H}) \mid \mathbf{G}, \vec{p}, \alpha)}{P(\text{Scalp}(\text{H}) \mid \mathbf{G}, \vec{p}, \alpha)} \quad (6.2)$$

Using the inequality from the hypothesis and Assumption 22, solve for α in (6.2). This gives $\alpha \leq 0.84507$. Since $\alpha_{\min} \leq \alpha$ we have $\alpha_{\min} \leq .84507$, so we can use Claim 5 to conclude that the likelihood ratio is less than 1. \square

Simplifying Assumption 7 (hypothesis of Claim 6). There exists $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ and $\vec{p} \in B$ such that

$$\frac{P(\text{News}(\text{H}) \mid \text{Beard}(\text{H}), \mathbf{G}, \vec{p}, \alpha)}{P(\text{News}(\text{H}) \mid \text{Scalp}(\text{H}), \mathbf{G}, \vec{p}, \alpha)} \leq 15$$

Goal Sentence 1. \langle the newspaper hair evidence is neutral or exculpatory \rangle

Proof. From Simplifying Assumption 7, Claim 6, and Assumption 21. \square

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 $p_2 > p_3 \geq p_4$

B_2 the subset of B where $p_3 > p_2 > p_4$

B_3 the subset of B where $p_3 > p_4 \geq p_2$

B_4 the subset of B where $p_4 > p_3 \geq p_2$

Claim 7. B is the disjoint union of B_1, B_2, B_3, B_4 .

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

$$\int_{\vec{p}=\langle p_1, p_2, p_3 \rangle \in B_1} t(p_1, p_2, p_3, 1-p_1-p_2-p_3) d\vec{p} = \int_{p_1=0}^{1/4} \int_{p_4=p_1}^{\frac{1-p_1}{3}} \int_{p_3=p_4}^{\frac{1-p_1-p_4}{2}} t(p_1, 1-p_1-p_3-p_4, p_3, p'_4) dp_1 dp_4 dp_3$$

$$\int_{\vec{p}=\langle p_1, p_2, p_3 \rangle \in B_2} t(p_1, p_2, p_3, 1-p_1-p_2-p_3) d\vec{p} = \int_{p_1=0}^{1/4} \int_{p_4=p_1}^{\frac{1-p_1}{3}} \int_{p_2=p_4}^{\frac{1-p_1-p_4}{2}} t(p_1, p_2, 1-p_1-p_2-p_4, p_4) dp_1 dp_4 dp_2$$

$$\int_{\vec{p}=\langle p_1, p_2, p_3 \rangle \in B_3} t(p_1, p_2, p_3, 1-p_1-p_2-p_3) d\vec{p} = \int_{p_1=0}^{1/4} \int_{p_2=p_1}^{\frac{1-p_1}{3}} \int_{p_4=p_2}^{\frac{1-p_1-p_2}{2}} t(p_1, p_2, 1-p_1-p_2-p_4, p_4) dp_1 dp_2 dp_4$$

$$\int_{\vec{p}=\langle p_1, p_2, p_3 \rangle \in B_4} t(p_1, p_2, p_3, 1-p_1-p_2-p_3) d\vec{p} = \int_{p_1=0}^{1/4} \int_{p_2=p_1}^{\frac{1-p_1}{3}} \int_{p_3=p_2}^{\frac{1-p_1-p_2}{2}} t(p_1, p_2, p_3, 1-p_1-p_2-p_3) dp_1 dp_2 dp_3$$

Claim 9. $\|B\| = 1/36$

Proof. The measure of B_j can be computed by standard means by substituting 1 in for $t(\dots)$ in the right side of the j -th equation of Claim 8. We find that $\|B_1\| = \|B_2\| = \|B_3\| = \|B_4\| = 1/144$. Hence $\|B\| = 1/36$ follows from Claim 7. \square

Claim 10. Simplified forms amenable to efficient computation:

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

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

Proof. The first equation follows easily from Simplifying Assumption 5 and Assumption 20. The second follows easily from Simplifying Assumption 5, Axioms 20 and 19, and Claim 4. \square

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

Claim 11.

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

$$\begin{aligned}
&= \frac{1}{(\alpha_{\max} - \alpha_{\min})\|\mathbf{B}\|} \sum_{i \in \{1,2,3,4\}} \int_{\alpha \in [\alpha_{\min}, \alpha_{\max}]} \int_{\vec{p} \in B_i} P(\text{Widths} = \vec{b} \mid \mathbf{G}, \vec{p}, \alpha) \\
&P(\text{Widths} = \vec{b} \mid \bar{\mathbf{G}}) = \int_{\vec{p} \in B} P(\text{Widths} = \vec{b} \mid \bar{\mathbf{G}}, \vec{p}) d_{\text{BParams}}(\vec{p} \mid \bar{\mathbf{G}}) \\
&= \frac{1}{\|\mathbf{B}\|} \sum_{i \in \{1,2,3,4\}} \int_{\vec{p} \in B_i} P(\text{Widths} = \vec{b} \mid \bar{\mathbf{G}}, \vec{p})
\end{aligned}$$

Proof. The first equation follows just from $\vec{p}, \alpha \mapsto P(\text{Widths} = \vec{b} \mid \mathbf{G}, \vec{p}, \alpha)$ being an integrable function and $d_{(\text{BParams}, \text{Mix})}(\vec{p}, \alpha \mid \mathbf{G})$ being the conditional density function for $\langle \text{Mix}, \text{BParams} \rangle$ given $\mathbf{G} = \text{true}$.

The second equation follows from Claim 7, Simplifying Assumptions 3 and 4, and the fact that $\vec{p}, \alpha \mapsto P(\text{Widths} = \vec{b} \mid \mathbf{G}, \vec{p}, \alpha)$ 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 (Definition 4) achieves its maximum of ≈ 1.27 when α_{\min} and α_{\max} are practically equal (unsurprising, as that allows the prosecution model to choose the best mixture parameter) and around .935; Plot 6.2 illustrates this, showing the likelihood ratio as a function of α_{\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 6.2, which shows the likelihood ratio as a function of α_{\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 α_{\min} , the quantity

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

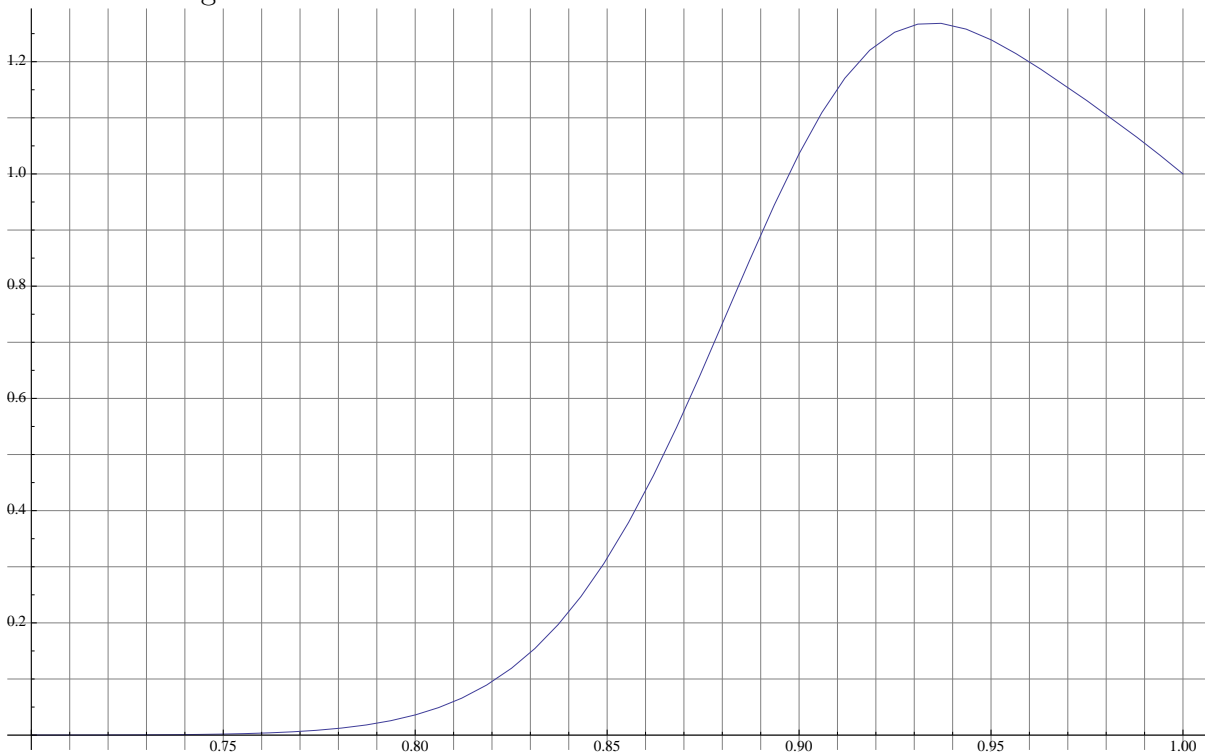
decreases as α_{\min} does. More precisely, Claim 5 follows from the following three propositions in Claim 12. The first has been tested using Mathematica's numerical integration; if it is false, it is unlikely to be false by a wide margin (i.e. taking a value slightly smaller than .849 should suffice). The remaining two have also not been proved, but one can gain good confidence in them by testing plots similar to Figure 6.2 for values of $\alpha_{\min} < .849$.

Proving Claim 12 or a slightly weaker version of it 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. Here, the notation $likelihood-ratio(\alpha_1, \alpha_2)$ is short for “the real number taken on by the defined term $likelihood-ratio$ when $\alpha_{\min} = \alpha_1$ and $\alpha_{\max} = \alpha_2$.”

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

Figure 6.1: Likelihood ratio as a function of α_{\min} when $\alpha_{\max} - \alpha_{\min} = 10^{-6}$, obtained by numerical integration.

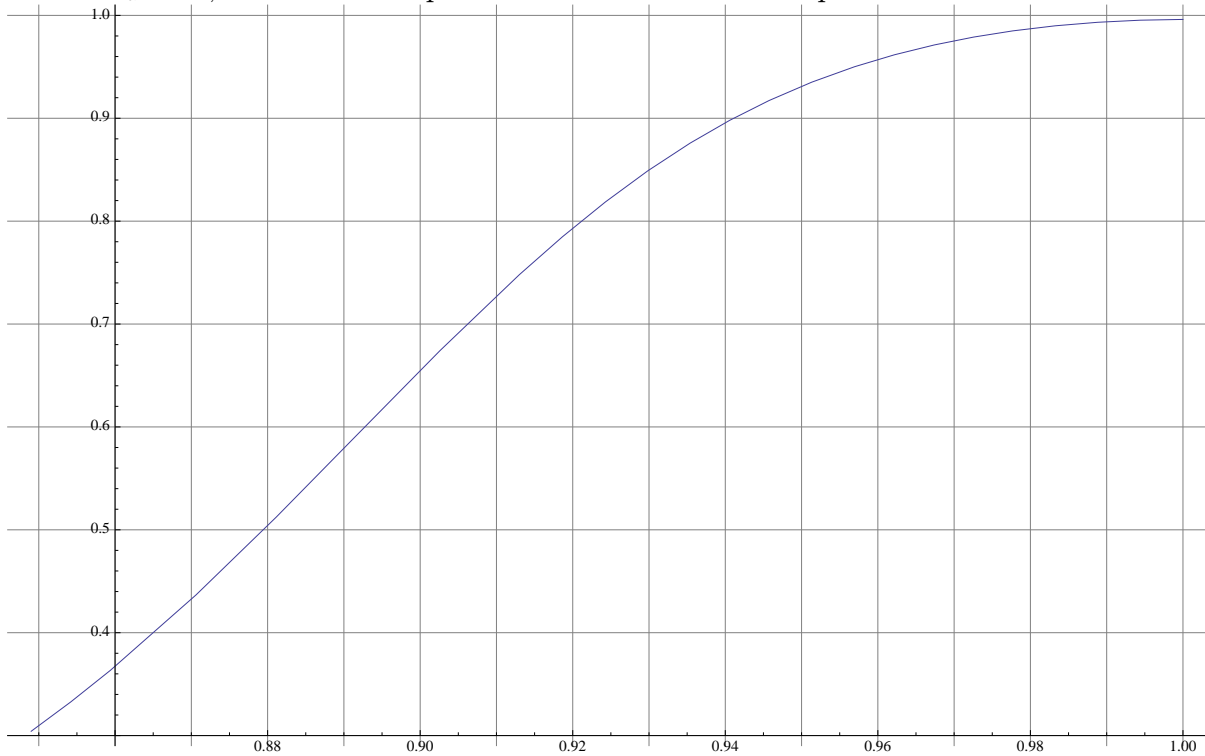


6.3 Criticism of argument

6.3.1 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 4**. In particular, the average value of the component of BParams for bin_1 ,

Figure 6.2: Likelihood ratio as a function of α_{\max} when $\alpha_{\min} = .849$, obtained by numerical integration. The shape of this plot is similar for smaller values of α_{\min} , being maximized when $\alpha_{\max} = 1$, which is what parts 2 and 3 of Claim 12 express.



the bin corresponding to the thinnest hairs, is 0.0625.⁸ It is best for the defense when the value of that component is 11/89, and best for the prosecution when it is 0, so the prosecution could reasonably insist that a prior is not fair unless the average is at most the mean of those two extremes, which is ≈ 0.0618 .

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

The defense can respond to the criticism, and I will explain how in Section 6.3.3. Doing so requires slightly strengthening the hypotheses of Claims 5 and 6.

6.3.2 Criticism 2

The second criticism says that 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

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

terms of the argument, the critic should reject the (implicit) axioms that constitute the types of the symbols `width` and `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, according to Section 2.2.1, one way to do this would be for the critic to suggest new definitions of `Bins`, `width`, and `Widths`. The critic can do this by suggesting new axioms (some of which are type constraints). Most importantly, the critic should suggest redefining the sort `Bins` as $\{\text{bin}_1, \dots, \text{bin}_5\}$, where `bin5` is a new constant. The results of that approach are discussed in the next section.

6.3.3 Response to criticisms

We can address both criticisms at once; if we introduce a fifth component of `BParams` corresponding to the interval $(187.5, \infty)$, and like the first component (probability `width` is in `bin1`) of `BParams` constrain it to be less than the middle three components (for `bin2`, `bin3`, `bin4`), then the average value of the `bin1` component of `BParams` goes down to $< .057$. We then need to slightly strengthen the hypotheses of the two main claims, changing the parameter `.85` in Claim 5 to `.835` and the parameter `15` in Claim 6 to `13.9`. Then, the proof works as before.

6.3.4 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*.⁹

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)

⁹Although 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.

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

Chapter 7

Example: Arguing that smoking causes cancer in 1950

In 1950 two landmark papers were published giving some of the first strong statistical evidence *in the English-speaking world*[SE05]¹ that tobacco smoking causes cancer, the first in the United States[WG85] and the second in England[DH50]. Yet it was not until 1965 that cigarette packages were required to have health warnings in the United States. Michael J. Thun, in his article *When truth is unwelcome: the first reports on smoking and lung cancer*, argued that 15 years was much too long given the strength of those studies:

In retrospect, the strength of the association in the two largest and most influential of these studies – by Ernest 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 *British 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 7,² too high to be dismissed as bias. [Thu05]

I will give part of an argument here that the health warnings policy was well-justified already in the early 1950s. The full argument involves introducing two more *candidate*

¹In Smith and Egger's short letter to the editors of the Bulletin of the World Health Organization[SE05], they give a very interesting account of how the history of this scientific progress is poorly known. In fact there were already *reviews* of the literature on the connection between smoking and lung cancer as early as 1929! Even the theory of second-hand smoking is at least as old as 1928.

²Note that these relative risk calculations treat the two studies separately, whereas both versions of the argument in this chapter use the earlier study to fit a model for the later study.

models (see below), the cigarette companies' *unknown genotype model* and the statistician R.A. Fisher's *soothing herb model*. How to refute those models is discussed in the next section. The part of the argument given here simply compares a weak version `dependModel` of the standard, causal model, 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(·, ·)` and Axiom 28) at predicting the true outcome (or an outcome near the true one) than the others. Moreover, suppose that M predicts 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 2.1)³, provided at least that one is firmly committed to certain ideals of persuasion.

Two versions of the argument are given, both dependent on mathematical claims that are unproved, but easily testable, and very likely easily resolvable by an appropriate expert (see footnote 1 on page 1 about unproved purely-mathematical claims). The version in Section 7.3 is simpler and more complete than the version in Section 7.2, but also weaker in that it uses a more idealized, less accurate model of the experiment.

7.1 Extensions and refinements of the argument

In the argument below, the causal scientific model, 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 experimental

³This example does not today meet the second criteria (contentiousness) that I listed there, but it did in the 1950s.

outcome distributions from more-qualitative assumptions. The reason is that those models are contrived to fit the data; they have outcome distributions similar to `dependModel`'s, seemingly (but not provably!) just 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. The easiest way to refute those models is to incorporate the data on female smoking and cancer, which neither model is able to explain without making them more elaborate.⁴ In fact, the argument could be strengthened in either or both of two ways: make models derive their experimental outcome distribution from qualitative assumptions, or make their experimental outcome distributions explain more data. I would advocate both.

Another, more technical and subtle way of improving the argument is to elaborate the definition of `Beats(M_1, M_2)` in such a way that, in effect, instantiations of the parameters of the two models M_1 and M_2 are only compared if they agree on the number of smokers in the British population. This would prevent one model from having an advantage over the other simply by having a better estimate of the total number of smokers (which, intuitively, we don't care about). Unfortunately, it would also very likely make the suitably-modified versions of Conjectures 1 and 2 harder to prove.

7.2 Proof with hypergeometric distributions contingent on an unproved mathematical claim

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

- CM : candidate models for the possible outcomes of the British study[DH50]. In the current version of this argument, a candidate model M is determined by `outcomeDistr(M, \cdot)` and `productWarning(M)`.

⁴Smoking 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.

- A : set of adult men living in the US at the time when the American study [WG85] was done.
- B : set of adult men living in England at the time when the British study was done.

Sharply-definable sorts (in $\mathcal{L}_{\text{rigid}}$)

- \mathbb{R} and \mathbb{N} - reals and natural numbers.
- \mathbb{R}^\perp and \mathbb{N}^\perp reals and naturals, but each with an extra element for “undefined”, to serve as the range of division and subtraction. Intuitively structures should make these be supersets of \mathbb{R} and \mathbb{N} , but technically all the sorts are disjoint. For readability I will not display the unary function symbols that are sometimes necessary to convert between the two sorts.
- $\text{FS}[\alpha]$ - finite subsets of (the interpretation of) the given sort α . This is a sort operator, i.e. a function from sorts to sorts.
- $\text{Fn}[\alpha, \beta]$ - the functions from α to β , another sort operator.
- Str - strings over the ASCII alphabet
- StudyOutcomes (informally a “subsort” or “subtype” of $\text{FS}[\mathbb{N}]$) - the set $\{620, \dots, 649\}$. Before the study is done, we don’t know how many of the people with lung cancer are smokers, i.e. $|\text{LC}_B^{\text{samp}} \cap \text{S}_B^{\text{samp}}|$ 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 $\text{LC}_B^{\text{samp}} \cap \text{S}_B^{\text{samp}}$) is $\{|\text{S}_B^{\text{samp}}| - |\overline{\text{LC}_B^{\text{samp}}}|, \dots, |\text{LC}_B^{\text{samp}}|\} = \{620, \dots, 649\}$.

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

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} : $\text{FS}[B]$ is a *hypothetical set*; the population that we imagine the British study samples were drawn from.
- LC_B^{pop} : $\text{FS}[B]$ is the set of people in B^{pop} with lung cancer.
- $\overline{\text{LC}_B^{\text{pop}}}$: $\text{FS}[B]$ is the set of people in B^{pop} without lung cancer.
- $\text{S}_{B,i}^{\text{pop}}$: $\text{FS}[B]$ is indepModel’s guess at the set of smokers in B^{pop} .
- $\text{S}_{B,d}^{\text{pop}}$: $\text{FS}[B]$ is dependModel’s guess at the set of smokers in B^{pop} .
- $A^{\text{samp}}, B^{\text{samp}}$: $\text{FS}[A]$ is the sample of patients used in the American (resp. British) study.
- $\text{LC}_A^{\text{samp}}, \text{LC}_B^{\text{samp}}$: $\text{FS}[A]$ is the set of people in A^{samp} (resp B^{samp}) who have lung cancer.

- $S_A^{\text{samp}}, S_B^{\text{samp}} : \text{FS}[A]$ is the set of smokers in A^{samp} (resp B^{samp}).
- $\text{outcomeDistr}(\cdot, \cdot) : \text{CM} \times \text{StudyOutcomes} \rightarrow \mathbb{R}$ is the given candidate model's distribution over StudyOutcomes .
- $\text{AllCM} : \text{FS}[\text{CM}]$ - the set of all candidate models. It should contain a candidate model from every interested party.

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

- $\text{StudyOutcomes} : \text{FS}[\mathbb{N}] := \{620, \dots, 649\}$. 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 $X \in \{\text{LC}_A^{\text{samp}}, S_A^{\text{samp}}\}$: $\bar{X} := A^{\text{samp}} \setminus X$
 - For each symbol $X \in \{\text{LC}_B^{\text{samp}}, S_B^{\text{samp}}\}$: $\bar{X} := B^{\text{samp}} \setminus X$
 - For each symbol $X \in \{\text{LC}_B^{\text{pop}}, S_{B,d}^{\text{pop}}, S_{B,i}^{\text{pop}}\}$: $\bar{X} := B^{\text{pop}} \setminus X$
- $\text{Pr}_{x \in U}(x \in V_1 \mid x \in V_2) : \text{FS}[\alpha] \times \text{FS}[\alpha] \times \text{FS}[\alpha] \rightarrow? \mathbb{R} := |V_1 \cap V_2 \cap U| / |V_2 \cap U|$
- For each $k \in \{0, 1, 2\}$:
 $\text{testInterval}_k : \text{FS}[\text{StudyOutcomes}] := \{ |S_B^{\text{samp}} \cap \text{LC}_B^{\text{samp}}| - k, \dots, |S_B^{\text{samp}} \cap \text{LC}_B^{\text{samp}}| + k \}$
- For each $k \in \{0, 1, 2\}$:

$$\text{test}_k(M) : \text{CM} \rightarrow \mathbb{R} := \sum_{x=\min(\text{testInterval}_k)}^{x=\max(\text{testInterval}_k)} \text{outcomeDistr}(M, x)$$

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

- $\text{Beats}(M_1:\text{CM}, M_2:\text{CM}) \leftrightarrow \bigwedge_{k \in \{0,1,2\}} \text{test}_k(M_1) > 1000 \cdot \text{test}_k(M_2)$. Model M_1 beats model M_2 if it assigns much higher probability to the true outcome $|S_B^{\text{samp}} \cap \text{LC}_B^{\text{samp}}|$, 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 .
- $\text{BeatsAll}(M_1:\text{CM}) \leftrightarrow \forall M_2:\text{CM}. (M_2 \in \text{AllCM} \wedge M_1 \neq M_2) \Rightarrow \text{Beats}(M_1, M_2)$ simply says that M_1 beats all the other models in AllCM .

Function symbols in $\mathcal{L}_{\text{rigid}}$

- $\{x, \dots, y\} : \mathbb{N} \times \mathbb{N} \rightarrow \text{FS}[\mathbb{N}]$ is the set of naturals from x to y inclusive, or the empty set if $x > y$.
- $+, \cdot : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$ are addition and multiplication for \mathbb{N} .

- $+, \cdot : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ addition and multiplication for \mathbb{R} . These symbols are distinct from the ones for \mathbb{N} , but display in the same way.
- $- : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}^\perp$ is subtraction, but undefined if the result is negative.
- $/ : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^\perp$ is division, undefined when the second argument is 0.
- **apply** : $\text{Fn}[\mathbb{N}, \mathbb{R}] \times \mathbb{N} \rightarrow \mathbb{R}$ is the application of a function object to an argument. There are versions of this symbol for a few other types as well.
- $\Sigma_{x=t_1}^{t_2} t_3(x) : \mathbb{N} \times \mathbb{N} \times \text{Fn}[\mathbb{N}, \mathbb{R}] \rightarrow \mathbb{R}$ is the usual summation binder symbol. The formal syntax is $\Sigma(t_1, t_2, \lambda x:\mathbb{N}.t_3)$, where “ $\lambda x:\mathbb{N}.t_3$ ” is actually just the name of a constant symbol of sort $\text{Fn}[\mathbb{N}, \mathbb{R}]$ that is implicitly defined in terms of the open term t_3 and the function symbol **apply**, but I’ve hidden those definitions from this writeup for readability.
- $\cap : \text{FS}[\alpha] \times \text{FS}[\alpha] \rightarrow \text{FS}[\alpha]$ is set intersection. As with the other function symbols in this list whose type is presented with a sort variable α , there are multiple distinct function symbols, for various instantiations of α , that each display as \cap . In the HTML presentation of interpreted formal proofs, one can disambiguate the symbol by hovering over it to see its type.
- $\setminus : \text{FS}[\alpha] \times \text{FS}[\alpha] \rightarrow \text{FS}[\alpha]$ is set difference.
- $|\cdot| : \text{FS}[\alpha] \rightarrow \mathbb{N}$ is the size of the given finite subset of (the interpretation of) α .
- $\binom{X}{k} : \text{FS}[\alpha] \times \mathbb{N} \rightarrow \text{FS}[\text{FS}[\alpha]]$ is the set of subsets of X of size k .
- $\min(\cdot), \max(\cdot) : \text{FS}[\mathbb{N}] \rightarrow \mathbb{N}^\perp$ are the minimum and maximum elements of a finite set of naturals. Undefined if the set is empty.
- $\text{hyper}(k, s, N, s') : \mathbb{N} \times \mathbb{N} \times \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}^\perp$ is the hypergeometric distribution (in the last argument; the other three arguments are parameters), defined when $s' \leq s \leq N, s \leq k \leq N$; 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 $\text{hyper}(k, s, N, s')$ is the probability that the resulting set contains exactly s' smokers.
- $\text{condHyper}(s_1, s_2, X_1, X_2, s'_1) : \mathbb{N} \times \mathbb{N} \times \text{FS}[B] \times \text{FS}[B] \times \mathbb{N} \rightarrow \mathbb{R}^\perp$ is a probability distribution (in the last argument; the other four arguments are parameters), defined when $s'_1 \leq s_1 \leq |S_B^{\text{samp}}| \leq N, s_1 \leq |X_1|, s_2 \leq |X_2|$. 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- $|LC_B^{\text{samp}}|$ subsets X'_1 of X_1 and X'_2 of X_2 . Then $\text{condHyper}(s_1, s_2, X_1, X_2, s'_1)$ is the conditional probability that X'_1 contains exactly s'_1 smokers, given that there are $|S_B^{\text{samp}}|$ smokers in $X'_1 \cup X'_2$.

Simplifying Assumption 8. We would change this to a normal Assumption if we included formalizations of Fisher’s and the tobacco companies’ models also (see section 7.1 above).

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

Quasi-Definition 3. Sizes of sets from the American study.

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

Quasi-Definition 4.

$\text{productWarning}(\text{dependModel}) =$ “*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 any person will get lung cancer.*”

$\text{productWarning}(\text{indepModel}) =$ “” (the empty string)

Quasi-Definition 5. 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 $\text{LC}_B^{\text{samp}} \cap \text{S}_B^{\text{samp}}$, given the sizes of $\text{LC}_B^{\text{samp}}$, $\overline{\text{LC}}_B^{\text{samp}}$, and S_B^{samp} . A model predicts the size well if its distribution over StudyOutcomes assigns high probability to $|\text{LC}_B^{\text{samp}} \cap \text{S}_B^{\text{samp}}|$ or some close number; this is formalized in the definition of $\text{Beats}(\cdot, \cdot)$.

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

Simplifying Assumption 9 (dependModel posits a hypergeometric distribution). Note that the values of the four parameters in

$\text{condHyper}(|\text{LC}_B^{\text{pop}} \cap \text{S}_{B,d}^{\text{pop}}|, |\overline{\text{LC}}_B^{\text{pop}} \cap \text{S}_{B,d}^{\text{pop}}|, \text{LC}_B^{\text{pop}}, \overline{\text{LC}}_B^{\text{pop}}, \cdot)$ are only bounded by the other axioms, especially Assumptions (23), (24), (25), and (26), with the latter two distinguishing dependModel ’s distribution from indepModel ’s. Still, this and Simplifying Assumption 10 are the weakest of the axioms with respect to the standards of accuracy that I strive for. Unlike the others, we cannot seriously claim that this axiom is literally true with respect to the informal intended semantics given by the language interpretation guide, 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 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.

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

Simplifying Assumption 10 (indepModel posits a hypergeometric distribution). Note that the values of the four parameters in $\text{condHyper}(|\text{LC}_B^{\text{pop}} \cap \text{S}_{B,i}^{\text{pop}}|, |\overline{\text{LC}}_B^{\text{pop}} \cap \text{S}_{B,i}^{\text{pop}}|, \text{LC}_B^{\text{pop}}, \overline{\text{LC}}_B^{\text{pop}}, \cdot)$ are only constrained by the other axioms, especially Assumptions (23), (24), and (27), with Assumption (27) distinguishing indepModel ’s distribution from dependModel ’s.

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

Assumption 23. 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 tripled from 1939 to 1950, then we should expect at most 6100 men with lung cancer.

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

Assumption 24. 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).

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

Assumption 25. Consider the ratio of probabilities of being a British nonsmoker given that you have lung cancer vs. given that you have a hospitalizable illness other than lung cancer. This axiom says that it is not much smaller than the corresponding ratio seen in the American study sample, specifically not more than 3 times smaller. 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 (26)), in effect positing that the correlation between smoking and lung cancer in the British population is identical to the correlation in the American sample. However, the evidence is so strongly in favor of `dependModel` that this much weaker assumption suffices:

$$\frac{\Pr_{x \in B^{\text{pop}}}(x \in \overline{S_{B,d}^{\text{pop}}} \mid x \in \overline{LC_B^{\text{pop}}})}{\Pr_{x \in B^{\text{pop}}}(x \in S_{B,d}^{\text{pop}} \mid x \in \overline{LC_B^{\text{pop}}})} \leq 3 \cdot \frac{\Pr_{x \in A^{\text{samp}}}(x \in \overline{S_A^{\text{samp}}} \mid x \in \overline{LC_A^{\text{samp}}})}{\Pr_{x \in A^{\text{samp}}}(x \in S_A^{\text{samp}} \mid x \in \overline{LC_A^{\text{samp}}})} \quad 5$$

Assumption 26. Same comment as in Assumption 25 applies here.

$$\Pr_{x \in B^{\text{pop}}}(x \in \overline{S_{B,d}^{\text{pop}}} \mid x \in \overline{LC_B^{\text{pop}}}) \geq 1/3 \cdot \Pr_{x \in A^{\text{samp}}}(x \in \overline{S_A^{\text{samp}}} \mid x \in \overline{LC_A^{\text{samp}}}) \quad 6$$

Assumption 27. 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.

$$\Pr_{x \in B^{\text{pop}}}(x \in S_{B,i}^{\text{pop}} \mid x \in \overline{LC_B^{\text{pop}}}) = \Pr_{x \in B^{\text{pop}}}(x \in S_{B,i}^{\text{pop}} \mid x \in LC_B^{\text{pop}})$$

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

Assumption 28. $\forall M:\text{CM}. \text{BeatsAll}(M) \Rightarrow \text{ShouldRequire}(\text{productWarning}(M))$

Claim 13.

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

equals

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

Proof. This is a standard definition of the conditional hypergeometric distribution. An informal proof is easy from the informal semantics given for `condHyper`($\cdot, \cdot, \cdot, \cdot, \cdot$). Note that we could alternatively have made `condHyper`($\cdot, \cdot, \cdot, \cdot, \cdot$) a defined function symbol. \square

The above axioms, together with some basic mathematical axioms, prove that for any setting of the free parameters $|LC_B^{\text{pop}}|, |\overline{LC_B^{\text{pop}}}|, |S_{B,i}^{\text{pop}} \cap LC_B^{\text{pop}}|, |S_{B,d}^{\text{pop}} \cap LC_B^{\text{pop}}|$, etc that

⁵ = $3(.0132231/.146154) \approx .27142$

⁶ = $(1/3).146154 = .048718$

obeys the constraints given by Axioms (23)-(27), 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:

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

7.3 Simpler, more-easily completable proof

We add a symbol for the binomial distribution.

$$\text{binDistr}_{\cdot, \cdot}(\cdot) : [0, 1] \times \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}^{\perp}$$

We may either give $\text{binDistr}_{\cdot, \cdot}(\cdot)$ a prose definition, and then state the next sentence as a Claim, or we could make $\text{binDistr}_{\cdot, \cdot}(\cdot)$ a defined function symbol defined by the next sentence. Either way is consistent with the definition of *interpreted formal proof*. In the proof from the previous section, the former option was taken. In this section I will leave it ambiguous.

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

We introduce a family of probability distributions that takes the place of $\text{condHyper}(\cdot, \cdot, \cdot, \cdot, \cdot)$.

We give it the following prose definition, and the later two axioms, Claims (14) and (15), are made only for the purpose of calculation.

- $\text{condBinom}(p_1, p_2, s'_1) : [0, 1] \times [0, 1] \times \mathbb{N} \rightarrow \mathbb{R}$ is a probability distribution (in the last argument; the other two arguments are parameters). Suppose we sample (with replacement) $|\text{LC}_B^{\text{samp}}|$ 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 $|\text{S}_B^{\text{samp}}|$.⁷

We also introduce three new constants $p_{\text{S}_d|\text{LC}}, p_{\text{S}_d|\overline{\text{LC}}}$, and $p_{\text{S}_d|*}$ of type $[0, 1]$. $p_{\text{S}_d|\text{LC}}$ and $p_{\text{S}_d|\overline{\text{LC}}}$ are **dependModel**'s estimates of the fraction of smokers in the lung cancer

⁷Note that this family of distributions is usually given with $|\text{LC}_B^{\text{samp}}|$ and $|\text{S}_B^{\text{samp}}|$ as parameters

population and in the population of people with conditions other than lung cancer. $p_{S_i|*}$ is `indepModel`'s estimate of the fraction of smokers in both populations.

We drop Simplifying Assumptions (9) and (10), replacing them with the following two:

Simplifying Assumption 11 (`dependModel` posits a binomial distribution). Note that the values of the two parameters (first two arguments) of `condBinom(·, ·, ·)` are only bounded by the axioms from the proof in the previous section and Assumptions (29), and (30), with the latter two distinguishing `dependModel`'s distribution from `indepModel`'s. The remainder of this paragraph (i.e. language interpretation guide entry) is essentially the same as in the description of Simplifying Assumption (9). This and Simplifying Assumption (12) are the weakest of the axioms with respect to the standards of accuracy 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 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.

$\forall s:\text{StudyOutcomes.}$

$$\text{outcomeDistr}(\text{dependModel}, s) = \text{condBinom}(p_{S_d|LC}, p_{S_d|\overline{LC}}, s)$$

Simplifying Assumption 12 (`indepModel` posits a binomial distribution).

$\forall s:\text{StudyOutcomes.}$

$$\text{outcomeDistr}(\text{indepModel}, s) = \text{condBinom}(p_{S_i|*}, p_{S_i|*}, s)$$

The next two axioms bound the frequencies mentioned in the previous two axioms. The description of Axiom (25) in the previous section has some motivation that applies here as well.

Assumption 29.

$$\frac{1}{2} \cdot \Pr_{x \in A^{\text{samp}}}(x \in \overline{S_A^{\text{samp}}} \mid x \in \overline{LC_A^{\text{samp}}}) \leq 1 - p_{S_d|\overline{LC}} \leq 2 \cdot \Pr_{x \in A^{\text{samp}}}(x \in \overline{S_A^{\text{samp}}} \mid x \in \overline{LC_A^{\text{samp}}})$$

Assumption 30.

$$\frac{1}{2} \cdot \Pr_{x \in A^{\text{samp}}}(x \in \overline{S_A^{\text{samp}}} \mid x \in LC_A^{\text{samp}}) \leq 1 - p_{S_d|LC} \leq 2 \cdot \Pr_{x \in A^{\text{samp}}}(x \in \overline{S_A^{\text{samp}}} \mid x \in LC_A^{\text{samp}})$$

The next two axioms tell us how to compute the distributions

Claim 14. For $n = |LC_B^{\text{samp}}|$ (and recall $|LC_B^{\text{samp}}| = |\overline{LC_B^{\text{samp}}}|$) have

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

equals

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

Proof. This is a standard symbolic expression for the conditional binomial distribution. Note that we could alternatively have made $\text{condBinom}(\cdot, \cdot, \cdot)$ a defined function symbol. \square

Claim 15. For $n = |LC_B^{\text{samp}}|$ (and recall $|LC_B^{\text{samp}}| = |\overline{LC_B^{\text{samp}}}|$) have

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

Proof. This is a well known fact, but I will provide the proof since it is short.

Let $n = |LC_B^{\text{samp}}|$ and $j = |SB|$. Referring to its prose definition, $\text{condBinom}(p, p, a)$ equals the probability of getting a successes when sampling n times from a distribution with success probability p , given that you've also sampled n times from another distribution with success probability p and the total of the two success counts is j . Equivalently, $\text{condBinom}(p, p, a)$ is

$$\frac{\text{Pr}(\text{get } a \text{ out of } n \text{ successes from first distribution and } j - a \text{ out of } n \text{ from second})}{\text{Pr}(\text{get total of } j \text{ successes out of } 2n)}$$

Because the two distributions are sampled from independently, that equals

$$\frac{\text{Pr}(\text{get } a \text{ out of } n \text{ successes from first}) \text{Pr}(\text{get } j - a \text{ out of } n \text{ from second})}{\text{Pr}(\text{get total of } j \text{ successes out of } 2n)}$$

i.e.

$$\frac{\text{binDistr}_{p, n}(a) \text{binDistr}_{p, n}(j - a)}{\text{binDistr}_{p, 2n}(j)}$$

Expanding the definition of binDistr that is:

$$\frac{\binom{n}{a} p^a (1 - p)^{n-a} \binom{n}{j-a} p^{j-a} (1 - p)^{n-j+a}}{\binom{2n}{j} p^j (1 - p)^{2n-j}} = \frac{\binom{n}{a} \binom{n}{j-a}}{\binom{2n}{j}}$$

□

The following Conjecture should be easier to prove than Conjecture 1, but it is still a tough (for a non-specialist) constrained nonlinear optimization problem.

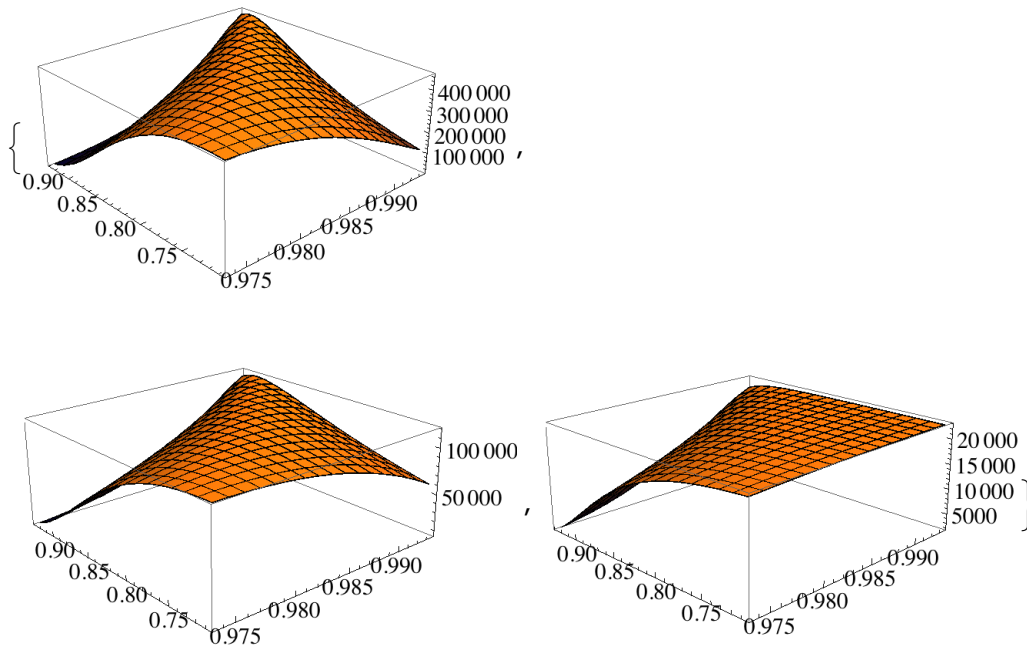
Conjecture 2.

$$\text{For } k \text{ in } \{0, 1, 2\} \text{ have } \text{test}_k(\text{dependModel}) > 2000 \cdot \text{test}_k(\text{indepModel})$$

Informal argument in support:

The independent variables model has no parameters, so $\text{test}_k(\text{indepModel})$ is a constant for each $k \in \{0, 1, 2\}$. For each $k \in \{0, 1, 2\}$, viewing the 3D plot of $\text{test}_k(\text{dependModel})$ as a function of the parameters $p_{S_d|LC}$ and $p_{S_d|\overline{LC}}$ (pictured below), if we make the very plausible assumption that there are no sharp extrema missed by the plot, then it is clear that within the range allowed by Axioms (29) and (30), the function is minimized at one of the corner points. In fact each of the three functions is minimized when $p_{S_d|LC}$ is minimal and $p_{S_d|\overline{LC}}$ is maximal. This is because the maximum likelihood model for the American data slightly overestimates the correlation between smoking and lung cancer in the British data. The minimums tell us that for any setting of the parameters of the dependModel that obeys the axioms, for $\text{test}_0(\cdot)$, $\text{test}_1(\cdot)$, and $\text{test}_2(\cdot)$ it gives probability more than 12, 000, 5, 000, and 2, 000 times higher, respectively, than indepModel.

Figure 7.1: Optimization problems for $k = 1, 2, 3$



Chapter 8

Example: Assisted suicide should be legalized in Canada

This is the most complex and ambitious of all the examples in this thesis. Not all of the many essentially-boolean-algebra lemmas have been formally verified using a first-order theorem prover, although any one of them can be checked easily by hand. Those remaining lemmas are stated as Claims.

This argument is meant to be read in a browser, and can be found at:

http://www.cs.toronto.edu/~wehr/thesis/assisted_suicide_msfol.html

I include an inferior static version here just in case you have a printed copy and you strongly prefer to read on paper.

Throughout this argument, S, NS, AS, PAS abbreviate 'suicide', 'no suicide', 'assisted suicide', 'physician-assisted suicide'

This is an argument for favouring the introduction of legislation that founds an administrative system (here broadly-defined) for granting Canadians access to physician-assisted suicide in limited cases. Before getting into the specifics, I want to discuss some important differences between this argument and (purely) natural language, non-deductive arguments also in favour of physician-assisted suicide (PAS).

The conclusion of this argument is weaker than the conclusions of the non-deductive arguments. In particular, the advocated (underspecified) PAS system is under-inclusive: it would not grant access to all people who I (or almost any proponent of physician-assisted suicide) think should have access. I am prepared to argue, informally, that this --and by "this" I mean under-inclusiveness in general, not the particular kind/degree of under-inclusive that the system involved here has-- that this is a necessary feature of any sufficiently-rigorous argument that does not make stronger assumptions than this one. More specifically, for any argument

- whose assumptions do not preclude the possibility of *regrettable* uses of physician-assisted suicide (see [◁regrettable PAS▷](#)), and
- that does not make strong assumptions along the lines of "regrettable assisted suicides aren't that bad", and
- whose assumptions do not guarantee that the advocated administrative system for PAS will have access to additional information about applicants,

there will be applicants who *should*, in a subjective moral sense according to proponents like me, have access to PAS, but who are indistinguishable from (hypothetical) applicants whose assisted suicide would be regrettable.

Even so, one could argue that this system is *especially* or *unnecessarily* under-inclusive, and that I won't dispute. This argument should be taken as a proof-of-concept. A more-serious attempt would have to involve much more research about this particular issue (e.g. data from the history of the systems in places like the Netherlands and Oregon) than I can afford to do for my thesis. A recent source of such information and much more is Quebec's [Committee report on dying with dignity](#).

At this point I could enumerate the ways that existing non-deductive/natural language arguments use stronger assumptions than those employed here. But here is a method that you can use to collect most of those ways yourself: for each Assumption and Simplifying Assumption, consider how one could use natural language to hide the details of the problem that the assumption addresses. For example, I will spend a surprising amount of time delineating the things that I

consider possibly-pertinent to the decision of whether or not to support the legislation. I have never seen this done in such a clear and explicit way in informal arguments; often the best one can hope for is that it can be unambiguously gleaned from reading nearly the *entire* text of the argument (which is book-length in some cases, e.g. the committee report linked to in the previous paragraph). Another example is the unusually-specific weighing of the (sometimes theoretical) pros and cons of such legislation; in this argument I go a level deeper than any natural language argument I have seen in the justification for the main required subjective proposition ✓[Lemma 74](#).

There may already exist a natural language argument about PAS that is every bit as fair and disciplined as the one I give here (the best *concise* argument I've read was written by the supreme court justices in the minority opinion for the Sue Rodriguez case). A similar thing can be said for any proposed prescription or standard; always there are actors in the scope of the prescription that have come to comply with it on their own. In the case of maximally fair/disciplined argumentation, like in the case of prescriptions for limiting pollution, or the protection of human rights, etc, those actors should be applauded, but their existence should not detract from the importance of the prescription unless they are sufficiently common (where "sufficiently common" is relative: we would need them to be very common in the case of pollution, or universal in the case of human rights).

Some features of the administrative system

First, a person must opt-in before being diagnosed with their terminal illness. The purpose of this feature is to lower the (already low) incidence of some forms of *difficult instances of regrettable assisted suicides* ([⟨difficult-regrettable PAS⟩](#)), which are regrettable assisted suicides in which

1. The applicant is correctly diagnosed; or the applicant is incorrectly diagnosed, but nonetheless suffers from an illness that is terminal without treatment, and all methods of diagnosis have been exhausted (see [⟨PAS with wrong and substandard diagnosis⟩](#)).
2. The person who uses the lethal drugs is the person who applied for them (see [⟨PAS of wrong person⟩](#)).
3. The applicant uses the drugs with the intention of dying.
4. The applicant was not actively coerced into applying for and/or taking the drugs (see [⟨coerced suicide⟩](#)).

Second, a person must take the drugs in the presence of their physician. This is not ideal, and is not required in Oregon, for example, but it is useful for justifying the assumption that with very high probability, the negations of items 2 and 3 above never happen.

Third, there is a fixed list of eligible terminal diseases, which includes only those for which it can be established with "high" probability that a person has the disease and will "very likely" not live for more than X months. Each of "high", "very likely" and X can differ for different patients; the first two only need to be high enough and the third small enough that the patient is unwilling to try to beat the doctors' reported odds. ⊕

Fourth, in some parts of Canada, prosecution for illegal assisted suicide is rare relative to its incidence. This must remain the case even after a PAS law is passed. This is to justify the assumption that [‹S, NS because of incr fear of prosecution›](#) happens to no one with high probability (in [Assumption 2](#)). In particular, a person answering a request to assist another person in dying, where the second person is ineligible for legal PAS, should not be, or perceive to be, at *increased* risk of prosecution, compared to the risk before the law is passed.

Fifth, to justify the assumption that [‹NS, NS & worse palliative care›](#) happens to no one with high probability (in [Assumption 2](#)), the government will monitor private donations towards palliative care and research on palliative care, and if the appropriately-adjusted funds per person drops below what it was before the PAS law was passed, the government will increase public funding to compensate.

Sixth, to support all of [Assumption 2](#) through [Assumption 6](#), since all can be broken by incompetence, the government will adopt one of the provisions from Quebec's [bill 52](#), which they summarize as follows:

A commission on end-of-life care is established under the name “Commission sur les soins de fin de vie”, as well as rules with respect to its composition and operations. The mandate of the Commission is to examine all matters relating to end-of-life care and to oversee the application of specific requirements relating to medical aid in dying.

High-level features of the argument

I refer to the things that the axioms assume are pertinent to comparing the two outcomes (legislation passed, or not) as **individual-future relations** (sort [IFR](#)). They are essentially just relations on the lives of people (sort [People](#)) in a given possible future (sort [Future](#)).

Classifying the differences of a person's life between two possible futures

We are going to be reasoning about the costs and benefits to people of the proposed legislation in terms of some of the qualitative effects, for each person, of "moving"

from a possible future without PAS (status-quo future) to one with PAS (assisted suicide future). To that end, we categorize people using **individual-future-difference relation** (sort Δ), which are relations about the features of their lives in a given pair of possible futures (sort $\diamond\text{FuturePair}$). Usually, a Δ is determined by a pair of IFR's; the people satisfying such a Δ satisfy the first IFR in the status-quo future and the second IFR in the assisted suicide future. Some Δ 's can't easily be represented as a product of two IFR's. An example is $\langle\text{NS}, \text{NS \& worse palliative care}\rangle$, which means:

The person's death is not a suicide in the status-quo future or the assisted suicide future, and the quality of palliative care available to them is worse in the assisted suicide future compared to in the status-quo future.

The problem is that *worse palliative care* depends on both of the futures. Such a Δ could be defined by a family of IFR-pairs, but that would appear to require quantifying the quality of palliative care, which is difficult to do right (in a way that is satisfactory to everyone), and even if done right introduces an unnecessary abstract entity into the argument (namely the partially-ordered set of quantities).

Distributions over possible futures

Our uncertainty about how the future will turn out, with or without new legislation for PAS, is modelled by a Bayesian probability distribution \mathcal{D} over pairs of possible futures, where the first (resp. second) element of a pair is a possible future in which the proposed legislation is not (resp. is) passed. To be accurate, the distribution is over equivalence classes of pairs of possible futures (sort $\diamond\text{FuturePairEC}$), where two pairs are in the same class if they agree on $\delta(\mathbf{p}, \cdot)$ for every individual future difference relation δ and every person \mathbf{p} (see definition of ec). That is just so that our interpretations of \mathcal{D} can be finite (since we define only finitely-many Δ 's), to assuage any discomfort about measures over possible futures.

Two typical futures

The language of Bayesian reasoning isn't used in informal debates about PAS, and it won't be used much here either. Instead, we will define two "typical" possible futures, according to whether or not the legislation is passed. In this introduction I will call them ASF (assisted suicide future) and SQF (status-quo future), though in the proof they are always paired together as \vec{F}_{typical} . **Assumption 1** is the axiom that says the goal $\langle\text{should pass}\rangle$ is implied by a statement about \vec{F}_{typical} , so it is the assumption a critic should reject if they have a problem with passing from Bayesian distributions to the typical future pair, or a more-fundamental problem with the use of a Bayesian distribution at all. For δ an individual-future-difference relation, and \vec{F} a pair of possible futures, let $\#(\delta, \vec{F})$ be the number of people \mathbf{p} such that $\delta(\mathbf{p})$

holds in \vec{F} (written $\delta(p, \vec{F})$ in the proof). Let $\text{Exp}\#(\delta, \vec{F})$ be the expected value of $\#(\delta, \vec{F})$ when \vec{F} is sampled from \mathcal{D} . Then in the definition of \vec{F}_{typical} , SQF and ASF are partially defined(\boxplus) in such a way that:

- extremely unlikely Δ 's don't happen to anyone.
- there are slightly-pessimistic counts (for proponent of PAS), relative to the expected values determined by \mathcal{D} , for Δ 's that are favourable for ASF.
- there are slightly-optimistic counts (for opponents of PAS), relative to the expected values determined by \mathcal{D} , for Δ 's that are favourable for SQF.

Argument in steps

Most of the main tasks of the argument are as follows:

1. Semantically define a number of IFR's (i.e. introduce them as primitive symbols, with language interpretation guide entries).
2. State uncontroversial axioms about relations between the IFR's.
3. Syntactically define a number of Δ 's in terms of IFR's, and semantically define a few more.
4. Derive some statements relating the Δ 's (using relations between IFR's), and assume some others as uncontroversial axioms.
5. Define the pair of typical futures $\vec{F}_{\text{typical}} = \langle \text{SQF}, \text{ASF} \rangle$ as discussed above.
6. Syntactically define sets of people P_1, \dots, P_n in terms of satisfying various Δ 's with respect to \vec{F}_{typical} .
7. Derive some subset, disjointness and emptiness relations between the P_i s (from relations between the Δ 's). Also make assumptions, some of them controversial, about relations between some of the other P_i s.
8. Make subjective, controversial assumptions of this form: "The change in moving from SQF to ASF in the quality-of-life for the people in P_i (considered as a group- not individually) is worse/better/approximately the same as the change for the people in P_j ."
9. State easily-agreed-upon axioms that suffice to derive, from the other axioms and lemmas, that the change in moving from SQF to ASF in the quality of life for the set of $\langle \text{all people} \rangle$ is positive overall.

Note that the form of the final conclusion in item 9 is the same as the form of the subjective assumptions in item 8. This may lead one to wonder what we gain from all this work. Here is part of the answer: the item 8 (and item 7) assumptions are much more specific than the item 9 conclusion, and because of this you learn a lot more about my opinions than you would from my simply asserting the item 9 conclusion. That is a good thing on its own, but it also makes the task of criticizing my opinions much more feasible.

[Hide reserved variable declarations](#)

Variables $\phi, \phi_1, \phi_2, \phi_3, \psi, \psi_1, \psi_2$ are reserved for sort IFR.

Variables $\delta, \delta_1, \delta_2, \delta_3, \delta_4$ are reserved for sort Δ .

Variable p is reserved for sort **People**.

Variables F, F_1, F_2, F_3, F_4 are reserved for sort $\diamond\text{Future}$.

Variables $\vec{F}, \vec{F}_1, \vec{F}_2$ are reserved for sort $\diamond\text{FuturePair}$.

Variables E, E_1, E_2 are reserved for sort $\diamond\text{FuturePairEC}$.

Variables P, Q, P_1, P_2, P_3, P_4 are reserved for sort $\text{FinSet}(\text{People})$.

Variables Y, Y_1, Y_2, Y_3 are reserved for sort $\text{MultiSet}(\Delta)$.

Hide sort operators

Sort op **FinSet** - Finite subsets of the given sort.

Sort op **MultiSet** - Finite multi-sets (unordered lists) of the given sort.

Sort op **Distr** \square

If the (interpretation of the) given sort S is finite, then it is the set of probability distributions over S . When S is not finite, you may define $\text{Distr}(S)$ however you like (e.g. arbitrarily make it a singleton set whose member is not in the interpretation of any other sort symbol.)

Show number predicate/function symbols

Hide sorts

Sort **People** \square

The set of residents of Canada. For concreteness, let's say the residents who are alive sometime during 1993 (the year of Sue Rodriguez's supreme court hearing) or later.

Sort $\diamond\text{Future}$ \square

Conceivable possible futures. This set is loosely defined. The main requirement, implicitly imposed by $\text{RelApp}_{\text{IFR}}$ and RelApp_{Δ} , is that each *element* is defined precisely enough that any **IFR** relation can be evaluated on it (and similarly for Δ relations).

Sort **IFR** \square

Individual-Future Relation. An element is a relation on $\text{People} \times \diamond\text{Future}$.

Sort $\diamond\text{FuturePair}$ - This is $\diamond\text{Future} \times \diamond\text{Future}$.

Sort Δ \square

Individual-future-difference relation. An element is just a relation on $\text{People} \times \diamond\text{FuturePair} = \text{People} \times \diamond\text{Future} \times \diamond\text{Future}$, which is accessed via RelApp_{Δ} .

Sort $\diamond\text{FuturePairEC}$ \square

Equivalence classes of $\diamond\text{Future}$ pairs. Effectively defined by **ec** and **rep**.

Sort $\text{FinSet}(\text{People})$

Sort $\text{MultiSet}(\Delta)$

Sort $\text{Distr}(\text{FPairEC})$

$\langle\text{should pass}\rangle : \rightarrow \mathbb{B}$ \square

True if you think that legislation should be passed that introduces an administrative system for assisted suicide that is compliant with the description

at the top of this argument.

Show definitions for pairing: (\cdot, \cdot) , $\cdot[1]$, $\cdot[2]$

Show definitions of \in , \notin , \emptyset

$\text{RelApp}_{\text{IFR}} : \text{IFR} \times \text{People} \times \diamond\text{Future} \rightarrow \mathbb{B}^{\boxminus}$

When ϕ is an IFR relation, p is from **People**, and F is from $\diamond\text{Future}$, then $\text{RelApp}_{\text{IFR}}(\phi, p, F)$ means ϕ is true for p in F . The second order syntax $\phi(p, F)$ is used to display $\text{RelApp}_{\text{IFR}}(\phi, p, F)$.

$\text{RelApp}_{\Delta} : \Delta \times \text{People} \times \diamond\text{FuturePair} \rightarrow \mathbb{B}^{\boxminus}$

When δ is a Δ relation, p is from **People**, and \vec{F} is from $\diamond\text{FuturePair}$, then $\text{RelApp}_{\Delta}(\delta, p, \vec{F})$ means δ is true for p in \vec{F} . The second order syntax $\delta(p, \vec{F})$ is used to display $\text{RelApp}_{\Delta}(\delta, p, \vec{F})$.

Hide set/relation equality axioms

Definitional Axiom 1: $\forall P_1, P_2. P_1 = P_2 \Leftrightarrow (\forall p. p \in P_1 \Leftrightarrow p \in P_2)$

Definitional Axiom 2: $\forall \phi_1, \phi_2. \phi_1 = \phi_2 \Leftrightarrow (\forall p. \forall F. \phi_1(p, F) \Leftrightarrow \phi_2(p, F))$

Definitional Axiom 3: $\forall \delta_1, \delta_2. \delta_1 = \delta_2 \Leftrightarrow (\forall p, \vec{F}. \delta_1(p, \vec{F}) \Leftrightarrow \delta_2(p, \vec{F}))$

$\Delta\text{of} : \text{People} \times \diamond\text{FuturePair} \rightarrow \Delta^{\boxminus}$

$\delta = \Delta\text{of}(p, \vec{F})$ is the unique most-specific/smallest $\delta : \Delta$ such that $\delta(p, \vec{F})$. For a given $\diamond\text{FuturePair} \vec{F}$, the differences (that you think are sufficiently-relevant to this debate) between a person p 's experiences in the two futures -directed differences, the change *from* the first element of \vec{F} to the second element of \vec{F} - are given by $\Delta\text{of}(p, \vec{F})$.

$\Delta\text{sOf} : \text{FinSet}(\text{People}) \times \diamond\text{FuturePair} \rightarrow \text{MultiSet}(\Delta)^{\boxminus}$

If P is a set of **People**, then $\Delta\text{sOf}(P, \vec{F})$ is the multiset of the same size $|P|$ obtained by applying $\Delta\text{sOf}(\cdot, \vec{F})$ to P .

$>_{\Delta} : \text{MultiSet}(\Delta) \times \text{MultiSet}(\Delta) \rightarrow \mathbb{B}^{\boxminus}$

A subjective comparison relation. If $Y_1 >_{\Delta} Y_2$ then you would rather have the life-experience changes Y_1 happen to $|Y_1|$ random people than have the life-experience changes Y_2 happen to $|Y_2|$ random people. Is a strict partial order.

$\approx : \text{MultiSet}(\Delta) \times \text{MultiSet}(\Delta) \rightarrow \mathbb{B}^{\boxminus}$

A subjective equivalence relation. If $Y_1 \approx Y_2$ then you are impartial or cannot decide whether you would rather have the life-experience changes Y_1 happen to $|Y_1|$ random people or the life-experience changes Y_2 happen to $|Y_2|$ random people.

Defn $\succeq_{\Delta} : \text{MultiSet}(\Delta) \times \text{MultiSet}(\Delta) \rightarrow \mathbb{B} - \forall Y_1, Y_2. Y_1 \succeq_{\Delta} Y_2 \Leftrightarrow (Y_1 >_{\Delta} Y_2 \vee Y_1 \approx Y_2)$

Defn $\sim \text{ec} : \diamond\text{FuturePair} \rightarrow \diamond\text{FuturePairEC} - \forall \vec{F}_1, \vec{F}_2. \text{ec}(\vec{F}_1) = \text{ec}(\vec{F}_2) \Leftrightarrow (\forall p, \delta. \delta(p, \vec{F}_1) \Leftrightarrow \delta(p, \vec{F}_2))$ - Equivalence class of the given $\diamond\text{FuturePair}$. See description of $\diamond\text{FuturePairEC}$.

Defn~ **rep** : $\diamond\text{FuturePairEC} \rightarrow \diamond\text{FuturePair} - \forall E. \text{ec}(\text{rep}(E)) = E \boxminus$

A representative for the given equivalence class. **rep**(E) can be any

$\vec{F} : \diamond\text{FuturePair}$ such that $\text{ec}(\vec{F}) = E$.

Defn **swap** : $\diamond\text{FuturePair} \rightarrow \diamond\text{FuturePair} - \forall \vec{F}. \text{swap}(\vec{F}) = (\vec{F}[2], \vec{F}[1])$

Defn~ **pair_{IFR}** : $\text{IFR} \times \text{IFR} \rightarrow \Delta - \forall p, \vec{F}, \phi, \psi. \text{pair}_{\text{IFR}}(\phi, \psi)(p, \vec{F}) \Leftrightarrow (\phi(p, \vec{F}[1]) \wedge \psi(p, \vec{F}[2])) \boxminus$

The Δ relation obtained in the natural way from two **IFR** relations; the first (resp. second) **IFR** relation is applied to the first (resp. second) element of the given $\diamond\text{FuturePair}$. Most elements of Δ can be defined in the way. An example of an exception is $\langle S, NS \ \& \ \text{worse palliative care} \rangle$, because ``worse palliative care'' depends on both elements of the $\diamond\text{FuturePair}$. Such a Δ relation could be defined by a family of **IFR**-pairs, although that would appear to require quantifying the quality of palliative care, which, if done right (in a way that is satisfactory to everyone), introduces an unnecessary abstract entity into the argument, namely the partially-ordered set of quantities.

Defn~ **peopleIn_{IFR}** : $\text{IFR} \times \diamond\text{Future} \rightarrow \text{FinSet}(\text{People}) - \forall \phi, F, p. p \in \text{peopleIn}_{\text{IFR}}(\phi, F) \Leftrightarrow \phi(p, F)$

Defn~ **peopleIn Δ** : $\Delta \times \diamond\text{FuturePair} \rightarrow \text{FinSet}(\text{People}) - \forall \delta, \vec{F}, p. p \in \text{peopleIn}_{\Delta}(\delta, \vec{F}) \Leftrightarrow \delta(p, \vec{F})$

Defn~ **0** : $\text{MultiSet}(\Delta) - \forall \vec{F}. 0 = \Delta\text{sOf}(\emptyset, \vec{F}) \boxminus$

Empty multiset that corresponds to a 'neutral' judgement; if $\Delta\text{sOf}(P, \vec{F}) = 0$, then for the set of people **P**, the pair of futures \vec{F} are, overall, of equal value with respect to the Δ relations in use.

Defn~ $\langle \text{true} \rangle$: $\text{IFR} - \forall F. \forall p. \langle \text{true} \rangle(p, F)$ - Trivial **IFR** such that $\langle \text{true} \rangle(p, F)$ for all p, F .

Show set function definitions

Show set predicate definitions

Show set facts

Hide IFR constants

$\langle S \rangle$: **IFR** - The person's death is a suicide.

Defn $\langle NS \rangle$: **IFR** := $\text{co}(\langle S \rangle)$ - The person's death is not a suicide.

Defn~ $\langle \text{PAS} \rangle$: **IFR** - $\langle \text{PAS} \rangle \subseteq \langle S \rangle$ - The person ends their life via physician-assisted suicide

Defn $\langle \text{non-PA suicide} \rangle$: **IFR** := $\langle S \rangle \setminus \langle \text{PAS} \rangle$ - The person's death is a non-physician-assisted suicide.

Defn~ $\langle \text{regrettable PAS} \rangle$: **IFR** - $\langle \text{regrettable PAS} \rangle \subseteq \langle \text{PAS} \rangle \boxminus$

A person **p** is in this **IFR** if they legally use physician-assisted suicide and if there is some information about **p**, which was unknown at the time when their application for PAS was approved, that, if it had been known, would

have caused a significant proportion (say, 20%) of people who would have supported p 's application to resolutely change their mind. Here "resolutely" means that no further information about p would again change the minds of those 20% of people. A more concise but possibly more vague definition would be to say that at least 20% of p 's supporters, if given "perfect information" about p , would change their minds.

Defn $\langle \text{non-regrettable PAS} \rangle : \text{IFR} := \langle \text{PAS} \rangle \setminus \langle \text{regrettable PAS} \rangle$

Defn~ $\langle \text{PAS with wrong and substandard diagnosis} \rangle : \text{IFR} - \langle \text{PAS with wrong and substandard diagnosis} \rangle \subseteq \langle \text{regrettable PAS} \rangle \square$

The person is incorrectly diagnosed with a soon-to-be-terminal condition when their real condition is presently treatable. For this argument, we will only allow PAS for conditions that can be diagnosed with very high accuracy. The other option one could take is to argue that a patient can still benefit from PAS even when they were wrongly diagnosed, provided at least that they had the best medical care available. The point is that, if all available evidence has been taken into account, including the probabilities of misleading and missing evidence, then the individual's decision indirectly expresses their utilities of living with suffering vs dying unnecessarily, and in that case we should use *their* utilities as opposed to an average.

Defn~ $\langle \text{coerced suicide} \rangle : \text{IFR} - \langle \text{coerced suicide} \rangle \subseteq \langle \text{S} \rangle \square$

The person was actively coerced into suicide (assisted or otherwise) by another person, where "actively coerced" means influenced by some means other than guilt, a sense of duty, etc. This is another logical possibility that is easy to "deal with" in this argument (where "deal with" does not mean prevent - see [Assumption 4](#)).

Defn $\langle \text{non-coerced suicide} \rangle : \text{IFR} := \langle \text{S} \rangle \setminus \langle \text{coerced suicide} \rangle$

Defn~ $\langle \text{PAS of wrong person} \rangle : \text{IFR} - \langle \text{PAS of wrong person} \rangle \subseteq \langle \text{regrettable PAS} \rangle \square$

This could *theoretically* happen in Oregon, for example, since the prescribed lethal drugs don't need to be taken in the presence of a physician. To simplify this argument, we posit a stricter system in which the prescribed drugs must be taken in the presence of a physician.

Defn $\langle \text{coerced PAS} \rangle : \text{IFR} := \langle \text{coerced suicide} \rangle \cap \langle \text{PAS} \rangle$

Definitional Axiom 4: $\langle \text{coerced PAS} \rangle \subseteq \langle \text{regrettable PAS} \rangle$

Defn $\langle \text{very bad PAS} \rangle : \text{IFR} := \langle \text{PAS of wrong person} \rangle \cup (\langle \text{coerced PAS} \rangle \cup \langle \text{PAS with wrong and substandard diagnosis} \rangle) \square$

Union of some relations that, when true, constitute major failures of the safeguards of the assisted suicide system.

Defn $\langle \text{non-coerced PAS} \rangle : \text{IFR} := \langle \text{PAS} \rangle \setminus \langle \text{coerced PAS} \rangle$

Defn $\langle \text{very bad non-coerced PAS} \rangle : \text{IFR} := \langle \text{very bad PAS} \rangle \cap \langle \text{non-coerced PAS} \rangle$

Defn $\langle \text{regrettable non-coerced PAS} \rangle : \text{IFR} := \langle \text{regrettable PAS} \rangle \cap \langle \text{non-coerced PAS} \rangle$

Defn $\langle \text{non-regrettable non-coerced PAS} \rangle : \text{IFR} := \langle \text{non-regrettable PAS} \rangle \cap \langle \text{non-coerced PAS} \rangle$

Defn $\langle \text{non-coerced non-PA suicide} \rangle : \text{IFR} := \langle \text{non-coerced suicide} \rangle \cap \langle \text{non-PA suicide} \rangle$

Defn $\langle \text{difficult-regrettable PAS} \rangle : \text{IFR} := \langle \text{regrettable PAS} \rangle \setminus \langle \text{very bad PAS} \rangle \boxminus$

The kinds of hypothetical regrettable uses of PAS that are hardest to prevent.

Defn $\langle \text{difficult-regrettable non-coerced PAS} \rangle : \text{IFR} := \langle \text{difficult-regrettable PAS} \rangle \cap \langle \text{non-coerced PAS} \rangle$

$\langle \text{NS \& worst-case denial of PAS} \rangle : \text{IFR} \boxminus$

These are the people who have a genuine self-centered desire for PAS and suffer most from not having the option of PAS. Includes at least each person who has an (eventually) physically and/or mentally painful condition and:

- Will not be physically capable of committing suicide without help, by the time they wish to. E.g. if their medical condition severely restricts their movement. And:

Doesn't know anyone who would be willing to assist them, or does know such a person but is unwilling to put them at risk of prosecution, or is unwilling to break the law.

Or:

- Would be physically capable, but
 - Can't afford to travel to a place where PAS is legal.
 And:
 - Is unwilling to end their life in a way that disfigures their body (e.g. for consideration of the person who finds them, or their family, or religious reasons).

$\langle \text{NS \& typical-case denial of PAS} \rangle : \text{IFR} \boxminus$

These are the people who have a genuine self-centered desire for PAS and *don't* suffer greatly from not having the option of PAS (i.e. not in $\langle \text{NS \& worst-case denial of PAS} \rangle$), but whose quality of life is still negatively affected.

$\langle \text{NS \& best-case denial of PAS} \rangle : \text{IFR} \boxminus$

These are the (possibly empty) set of people who have a genuine self-centered desire for PAS but nonetheless *benefit*, overall, from not having the option of PAS.

Defn $\sim \langle \text{NS \& desire for PAS} \rangle : \text{IFR} - \text{DisjointUnion}_3(\langle \text{NS \& desire for PAS} \rangle, \langle \text{NS \& worst-case denial of PAS} \rangle, \langle \text{NS \& typical-case denial of PAS} \rangle, \langle \text{NS \& best-case denial of PAS} \rangle) \wedge \langle \text{NS \& desire for PAS} \rangle \subseteq \langle \text{NS} \rangle - \text{Person wanted PAS but didn't receive it (for any reason).}$

Defn $\langle \text{NS \& no desire for PAS} \rangle : \text{IFR} := \langle \text{NS} \rangle \setminus \langle \text{NS \& desire for PAS} \rangle$

Hide Δ constants

Show trivial Δ definitions

$\langle \text{true, NS \& worse palliative care} \rangle : \Delta \boxminus$

A genuine worry for some opponents of PAS is that a PAS policy could lead to decreased funding for palliative care. There are two ways one can address this concern: (1) use data from jurisdictions that have PAS to argue that it

hasn't happened there, and postulate that Canada will be the same. (2) take extra, active measures to defend against this happening. I have only touched on (2); see note about this in introduction (starts "Fifth, to justify the assumption...")

Defn $\langle S, NS \ \& \ \text{worse palliative care} \rangle : \Delta := \langle \text{true}, NS \ \& \ \text{worse palliative care} \rangle \cap \langle S, NS \rangle$

Defn $\langle NS, NS \ \& \ \text{worse palliative care} \rangle : \Delta := \langle \text{true}, NS \ \& \ \text{worse palliative care} \rangle \cap \langle NS, NS \rangle \boxminus$

The person's death is not a suicide in the status-quo future or the assisted suicide future, and the quality of palliative care available to them is worse in the assisted suicide future compared to in the status-quo future.

Defn~ $\langle S, NS \ \text{because of incr fear of prosecution} \rangle : \Delta - \langle S, NS \ \text{because of incr fear of prosecution} \rangle \subseteq \langle S, NS \rangle \boxminus$

There is a logical possibility that a person will choose to end their life via illegal non-physician-assisted suicide if legal PAS is not an option for anyone, but will choose *not* to if legal PAS is an option for some people but not for them. If the difference is because they have an increased fear that whoever helps them to die will be prosecuted, then they would be worse off in the assisted-suicide future (with a new law passed) than they would be in the status-quo future. We will specify that any PAS law that this argument endorses does not lead to increased prosecution for illegal assisted suicide.

$\langle NS, NS \ \text{because of incr fear of prosecution} \rangle : \Delta$

Definitional Axiom 5: $\langle NS, NS \ \text{because of incr fear of prosecution} \rangle \subseteq \langle NS, NS \rangle$

Defn $\langle \text{true}, NS \ \text{because of incr fear of prosecution} \rangle : \Delta := \langle NS, NS \ \text{because of incr fear of prosecution} \rangle \cup \langle S, NS \ \text{because of incr fear of prosecution} \rangle$

Definitional Axiom 6: $\text{Disjoint}(\langle NS, NS \ \text{because of incr fear of prosecution} \rangle, \langle S, NS \ \text{because of incr fear of prosecution} \rangle)$

Defn $\langle NS, NS \ \& \ \text{same-or-better palliative care} \rangle : \Delta := \langle NS, NS \rangle \setminus \langle NS, NS \ \& \ \text{worse palliative care} \rangle$

Defn $\langle S, NS \ \& \ \text{same-or-better palliative care} \ \& \ \text{no incr fear of prosecution} \rangle : \Delta := (\langle S, NS \rangle \setminus \langle S, NS \ \text{because of incr fear of prosecution} \rangle) \setminus \langle S, NS \ \& \ \text{worse palliative care} \rangle$

$\langle NS, NS \ \& \ \text{same-or-better palliative care} \ \& \ \text{exacerbated guilt} \rangle : \Delta \boxminus$

A genuine worry for some opponents of assisted suicide is that, with physician-assisted suicide available, we will see a decrease in the quality of life of people who are eligible for it but do not want it, due to, for example, feelings of being a burden on their family and friends. We do not try to design an assisted suicide system that prevents this; instead we acknowledge it and postulate that it is compensated for by the benefits of an assisted suicide system.

$\langle NS, NS \ \& \ \text{same-or-better palliative care} \ \& \ \text{no exacerbated guilt} \rangle : \Delta$

Definitional Axiom 7: $\text{DisjointUnion}(\langle NS, NS \ \& \ \text{same-or-better palliative care} \rangle, \langle NS, NS \ \& \ \text{same-or-better palliative care} \ \& \ \text{no exacerbated guilt} \rangle, \langle NS, NS \ \& \ \text{same-or-better palliative care} \ \& \ \text{exacerbated guilt} \rangle)$

Defn **EasierCases** : $\Delta := (\langle \text{true}, \text{NS \& worse palliative care} \rangle \cup \langle \text{true}, \text{NS because of incr fear of prosecution} \rangle) \cup (\text{pair}_{\text{IFR}}(\langle \text{true} \rangle, \langle \text{very bad PAS} \rangle) \cup \langle \text{S}, \text{S} \rangle)$
 Defn **$\langle \text{NS}, \text{coerced suicide} \rangle$** : $\Delta := \text{pair}_{\text{IFR}}(\langle \text{NS} \rangle, \langle \text{coerced suicide} \rangle) \setminus \text{EasierCases}$
 Defn **$\langle \text{coerced suicide}, \text{NS} \rangle$** : $\Delta := \text{pair}_{\text{IFR}}(\langle \text{coerced suicide} \rangle, \langle \text{NS} \rangle) \setminus \text{EasierCases}$
 Defn **$\langle \text{NS}, \text{non-PA suicide \& non-coerced} \rangle$** : $\Delta := \text{pair}_{\text{IFR}}(\langle \text{NS} \rangle, \langle \text{non-PA suicide} \rangle \cap \langle \text{non-coerced suicide} \rangle) \setminus \text{EasierCases}$
 Defn **$\langle \text{non-PA suicide \& non-coerced}, \text{NS} \rangle$** : $\Delta := \text{pair}_{\text{IFR}}(\langle \text{non-PA suicide} \rangle \cap \langle \text{non-coerced suicide} \rangle, \langle \text{NS} \rangle) \setminus \text{EasierCases}$
 Defn **$\langle \text{NS}, \text{difficult-regrettable PAS} \rangle$** : $\Delta := \text{pair}_{\text{IFR}}(\langle \text{NS} \rangle, \langle \text{difficult-regrettable PAS} \rangle) \setminus \text{EasierCases}$
 Defn **$\langle \text{NS \& typical-case denial of PAS}, \text{non-regrettable PAS} \rangle$** : $\Delta := \text{pair}_{\text{IFR}}(\langle \text{NS \& typical-case denial of PAS} \rangle, \langle \text{non-regrettable PAS} \rangle) \setminus \text{EasierCases}$
 Defn **$\langle \text{NS \& worst-case denial of PAS}, \text{non-regrettable PAS} \rangle$** : $\Delta := \text{pair}_{\text{IFR}}(\langle \text{NS \& worst-case denial of PAS} \rangle, \langle \text{non-regrettable PAS} \rangle) \setminus \text{EasierCases}$
 Defn **$\langle \text{S}, \text{NS \& no incr fear of prosecution} \rangle$** : $\Delta := \langle \text{S}, \text{NS} \rangle \setminus \langle \text{S}, \text{NS because of incr fear of prosecution} \rangle$
 Defn **$\langle \text{S}, \text{NS \& worse palliative care \& no incr fear of prosecution} \rangle$** : $\Delta := \langle \text{S}, \text{NS \& worse palliative care} \rangle \cap \langle \text{S}, \text{NS \& no incr fear of prosecution} \rangle$

Note that, at the current level of detail of this proof, the next two symbols are only used in the semantic descriptions of other symbols - not in any axioms.

\mathcal{D} : $\text{Distr}(\diamond \text{FuturePairEC})$ - A finite, subjective Bayesian probability distribution over $\diamond \text{FuturePairEC}$

Exp# : $\text{Distr}(\diamond \text{FuturePairEC}) \times \Delta \rightarrow \mathbb{R}^{\geq 0} \boxminus$

$\text{Exp#}(D, \delta)$ is the expected value of $|\text{peopleIn}_{\Delta}(\delta, \text{rep}(E))|$ when $\diamond \text{FuturePairEC}$ E is chosen at random from D

Defn \sim **$\langle \text{favourable for status quo} \rangle$** : $\Delta \rightarrow \mathbb{B} - \forall \delta. \langle \text{favourable for status quo} \rangle(\delta) \Leftrightarrow 0 >_{\Delta} \Delta \text{sOf}(\text{peopleIn}_{\Delta}(\delta, \vec{F}_{\text{typical}}), \vec{F}_{\text{typical}}) \boxminus$

$\langle \text{favourable for status quo} \rangle(\delta)$ means that $\delta(\cdot, \vec{F}_{\text{typical}})$ represents a negative change (or equivalently, $\delta(\cdot, \text{swap}(\vec{F}_{\text{typical}}))$ represents a positive change). This is more easily conveyed by examples: $\langle \text{favourable for status quo} \rangle(\langle \text{NS}, \text{coerced suicide} \rangle)$ holds because if $\langle \text{NS}, \text{coerced suicide} \rangle(p, \vec{F}_{\text{typical}})$ then clearly something gets better for p when moving from the assisted suicide future (the second of the pair) to the status quo future (or worse when moving in the opposite direction).

\vec{F}_{typical} : $\diamond \text{FuturePair} \boxminus$

A particular future pair whose equivalence class is "typical", in a certain sense, with respect to \mathcal{D} . At a high level, we assume

- Extremely unlikely (w.r.t. \mathcal{D}) Δ s don't happen to anyone in \vec{F}_{typical} .
- The count of people who δ happens to is roughly the expected value of the corresponding counting random variable of \mathcal{D} . We round up for the Δ s that are favourable for the status-quo (SQ) futures, and down for the Δ s that are neutral or favourable for the assisted suicide futures.

More precisely, there is a constant $\varepsilon: \mathbb{R}^{\geq 0}$, which is a parameter of this argument, such that:

- $|\text{peopleIn}_{\Delta}(\delta, \vec{F}_{\text{typical}})| = 0$ for all δ such that $\text{Exp}\#(\delta, \mathcal{D}) < \varepsilon$.
- $|\text{peopleIn}_{\Delta}(\delta, \vec{F}_{\text{typical}})| = \text{ceil}(x)$ for all δ such that $\text{Exp}\#(\delta, \mathcal{D}) = x \geq \varepsilon$ and $\langle \text{favourable for status quo} \rangle(\delta)$.
- $|\text{peopleIn}_{\Delta}(\delta, \vec{F}_{\text{typical}})| = \text{floor}(x)$ for all δ such that $\text{Exp}\#(\delta, \mathcal{D}) = x \geq \varepsilon$ and $\neg \langle \text{favourable for status quo} \rangle(\delta)$.

Defn $\sim \langle \text{all people} \rangle : \text{FinSet}(\text{People}) - \forall p. p \in \langle \text{all people} \rangle$

Assumption 1: $\Delta \text{sOf}(\langle \text{all people} \rangle, \vec{F}_{\text{typical}}) >_{\Delta} 0 \Rightarrow \langle \text{should pass} \rangle$

\exists Goal: $\langle \text{should pass} \rangle$

Defn $\sim \text{peopleIn}_{\text{IFR}}^{1,*} : \text{IFR} \rightarrow \text{FinSet}(\text{People}) - \forall \phi. \forall p. p \in \text{peopleIn}_{\text{IFR}}^{1,*}(\phi) \Leftrightarrow \phi(p, \vec{F}_{\text{typical}}[1])$

Defn $\sim \text{peopleIn}_{\text{IFR}}^{2,*} : \text{IFR} \rightarrow \text{FinSet}(\text{People}) - \forall \phi. \forall p. p \in \text{peopleIn}_{\text{IFR}}^{2,*}(\phi) \Leftrightarrow \phi(p, \vec{F}_{\text{typical}}[2])$

Defn $\text{peopleIn}_{\Delta} : \Delta \rightarrow \text{FinSet}(\text{People}) - \forall \delta. \text{peopleIn}_{\Delta}(\delta) = \text{peopleIn}_{\Delta}(\delta, \vec{F}_{\text{typical}})$

Hide $>_{\Delta}$ axioms

Show " \approx is an equivalence relation"

Show " $>_{\Delta}$ is a strict partial order."

Definitional Axiom 14: $\forall P_1, P_2, \vec{F}. \text{Disjoint}(P_1, P_2) \wedge \Delta \text{sOf}(P_1, \vec{F}) \approx 0 \wedge \Delta \text{sOf}(P_2, \vec{F}) \approx 0 \Rightarrow \Delta \text{sOf}(P_1 \cup P_2, \vec{F}) \approx 0$

Definitional Axiom 15: $\forall P_1, P_2. \forall \vec{F}. \text{Disjoint}(P_1, P_2) \wedge \Delta \text{sOf}(P_1, \vec{F}) \approx_{\Delta} 0 \wedge \Delta \text{sOf}(P_2, \vec{F}) \approx_{\Delta} 0 \Rightarrow \Delta \text{sOf}(P_1 \cup P_2, \vec{F}) \approx_{\Delta} 0$

Definitional Axiom 16: $\forall P_1, P_2. \forall \vec{F}. \text{Disjoint}(P_1, P_2) \wedge \Delta \text{sOf}(P_1, \vec{F}) \approx_{\Delta} 0 \wedge \Delta \text{sOf}(P_2, \vec{F}) >_{\Delta} 0 \Rightarrow \Delta \text{sOf}(P_1 \cup P_2, \vec{F}) >_{\Delta} 0$

Definitional Axiom 17: $\forall P_1, P_2. \forall \vec{F}. \text{Disjoint}(P_1, P_2) \wedge \Delta \text{sOf}(P_1, \vec{F}) \approx \Delta \text{sOf}(P_2, \text{swap}(\vec{F})) \Rightarrow \Delta \text{sOf}(P_1 \cup P_2, \vec{F}) \approx 0$

Definitional Axiom 18: $\forall P_1, P_2. \forall \vec{F}. \text{Disjoint}(P_1, P_2) \wedge \Delta \text{sOf}(P_1, \vec{F}) \approx_{\Delta} \Delta \text{sOf}(P_2, \text{swap}(\vec{F})) \wedge \Delta \text{sOf}(P_1, \vec{F}) \approx_{\Delta} 0 \Rightarrow \Delta \text{sOf}(P_1 \cup P_2, \vec{F}) \approx_{\Delta} 0$

Definitional Axiom 19: $\forall P_1, P_2. \forall \vec{F}. \text{Disjoint}(P_1, P_2) \wedge \Delta\text{sOf}(P_1, \vec{F}) >_{\Delta} \Delta\text{sOf}(P_2, \text{swap}(\vec{F})) \wedge \Delta\text{sOf}(P_1, \vec{F}) \approx_{\Delta} 0 \Rightarrow \Delta\text{sOf}(P_1 \cup P_2, \vec{F}) >_{\Delta} 0$

Assumption 2: For each δ in $\{\text{pair}_{\text{IFR}}(\langle \text{true} \rangle, \langle \text{PAS of wrong person} \rangle), \text{pair}_{\text{IFR}}(\langle \text{true} \rangle, \langle \text{PAS with wrong and substandard diagnosis} \rangle), \langle \text{S, NS because of incr fear of prosecution} \rangle, \langle \text{NS, NS \& worse palliative care} \rangle\}$: $\text{peopleIn}_{\Delta}(\delta) = \emptyset$
 \square

The proposed safeguards for the implementation make these Δ s so unlikely that they do not apply to anyone in \vec{F}_{typical} .

Assumption 3: For each δ in $\{\text{pair}_{\text{IFR}}(\langle \text{NS \& no desire for PAS} \rangle, \langle \text{non-regrettable PAS} \rangle), \langle \text{NS, NS \& same-or-better palliative care \& no exacerbated guilt} \rangle, \langle \text{S, S} \rangle, \langle \text{S, NS \& same-or-better palliative care \& no incr fear of prosecution} \rangle\}$: $\Delta\text{sOf}(\text{peopleIn}_{\Delta}(\delta), \vec{F}_{\text{typical}}) \approx_{\Delta} 0$
 \square

The quality of life for these people is not significantly different between the two futures. We make the superficially-weaker assumption that their quality of life is not worse in the assisted-suicide future than it is in the status quo future.

Assumption 4: $\Delta\text{sOf}(\text{peopleIn}_{\Delta}(\langle \text{NS, coerced suicide} \rangle), \vec{F}_{\text{typical}}) \approx \Delta\text{sOf}(\text{peopleIn}_{\Delta}(\langle \text{coerced suicide, NS} \rangle), \text{swap}(\vec{F}_{\text{typical}}))$
 \square

Let P_1 be the people whose death is not a suicide in SQF and whose death is a coerced suicide in ASF. Let P_2 be the people whose death is a coerced suicide in SQF and not a suicide in ASF. Note that P_1 and P_2 are disjoint. When it comes to comparing SQF and ASF, the cost for the people in P_1 of moving from SQF to ASF is approximately equal to the cost for the people in P_2 of moving from ASF to SQF. In a more-detailed version of this argument, this assumption would be derived (by a single axiom) from two simpler ones:

- The size of P_1 is approximately equal to the size of P_2 . This requires that passing of PAS legislation eliminates as many coerced suicides as it introduces (P_2 are eliminated, P_1 are introduced).
- You believe coerced suicide in ASF and coerced suicide in SQF are equally-bad fates.

Assumption 5: $\Delta\text{sOf}(\text{peopleIn}_{\Delta}(\langle \text{NS \& typical-case denial of PAS, non-regrettable PAS} \rangle), \vec{F}_{\text{typical}}) \approx_{\Delta} \Delta\text{sOf}(\text{peopleIn}_{\Delta}(\langle \text{NS, NS \& same-or-better palliative care \& exacerbated guilt} \rangle), \text{swap}(\vec{F}_{\text{typical}}))$
 \square

Let P_1 be the people who satisfy $\langle \text{NS \& typical-case denial of PAS} \rangle$ in SQF and $\langle \text{non-regrettable PAS} \rangle$ in ASF. Let P_2 be the people whose death is not a suicide in SQF who, in ASF, still do not die by suicide, and have equally-good access to palliative care, but also are tormented with guilt for not choosing PAS (and there was no such guilt for them in SQF, because PAS was illegal). When it comes to comparing SQF and ASF, the cost to the people in P_1 of moving from ASF to SQF is at least as great as the cost to the people in

P_2 of moving from SQF to ASF. In a more-detailed version of this argument, this assumption would be derived (by a single axiom) from two simpler ones:

- For some fraction $n/m \geq 1$ (say, $n=10$ and $m=1$), $(n/m)|P_1| < |P_2|$. That is, you grant that P_2 is larger than P_1 , but you posit that it is no more than (n/m) times larger.
- You believe it is at least as bad to move m people who satisfy $\langle \text{NS \& typical-case denial of PAS, non-regrettable PAS} \rangle$ from ASF to SQF as it is to move n people who satisfy $\langle \text{NS, NS \& same-or-better palliative care \& exacerbated guilt} \rangle$ from SQF to ASF.

Assumption 6: $\Delta \text{Of}(\text{peopleIn}_{\Delta}(\langle \text{NS \& worst-case denial of PAS, non-regrettable PAS} \rangle), \vec{F}_{\text{typical}}) >_{\Delta} \Delta \text{Of}(\text{peopleIn}_{\Delta}(\langle \text{NS, difficult-regrettable PAS} \rangle), \text{swap}(\vec{F}_{\text{typical}})) \boxminus$

This assumption is similar to [Assumption 5](#) in form (except for using $>_{\Delta}$ instead of \succeq_{Δ}), but it concerns more-severe kinds of changes. This assumption is often the main *explicit* disagreement between non-religious proponents of PAS and (self-reported) non-religious opponents of PAS. So please read the informal description of [Assumption 5](#) first. Like [Assumption 5](#), in a more-detailed version of this argument, this assumption would be derived (by a single axiom) from two simpler ones, which are of the form in [Assumption 5](#) *except* that one might argue the fraction n/m is < 1 if the safeguards of the PAS system are very good.

Hide definitions of sets of people corresponding to previous 5 assumptions.

Defn **People₁** : $\text{FinSet}(\text{People}) := \text{peopleIn}_{\Delta}(\text{pair}_{\text{IFR}}(\langle \text{true} \rangle, \langle \text{PAS with wrong and substandard diagnosis} \rangle)) \cup (\text{peopleIn}_{\Delta}(\text{pair}_{\text{IFR}}(\langle \text{true} \rangle, \langle \text{PAS of wrong person} \rangle)) \cup (\text{peopleIn}_{\Delta}(\langle \text{NS, NS \& worse palliative care} \rangle) \cup \text{peopleIn}_{\Delta}(\langle \text{S, NS because of incr fear of prosecution} \rangle)))$

Defn **People₂** : $\text{FinSet}(\text{People}) := ((\text{peopleIn}_{\Delta}(\langle \text{S, S} \rangle) \cup \text{peopleIn}_{\Delta}(\langle \text{S, NS \& same-or-better palliative care \& no incr fear of prosecution} \rangle)) \cup \text{peopleIn}_{\Delta}(\langle \text{NS, NS \& same-or-better palliative care \& no exacerbated guilt} \rangle) \cup \text{peopleIn}_{\Delta}(\text{pair}_{\text{IFR}}(\langle \text{NS \& no desire for PAS} \rangle, \langle \text{non-regrettable PAS} \rangle))$

Defn **People₃** : $\text{FinSet}(\text{People}) := \text{peopleIn}_{\Delta}(\langle \text{NS, coerced suicide} \rangle) \cup \text{peopleIn}_{\Delta}(\langle \text{coerced suicide, NS} \rangle)$

Defn **People₄** : $\text{FinSet}(\text{People}) := \text{peopleIn}_{\Delta}(\langle \text{NS, NS \& same-or-better palliative care \& exacerbated guilt} \rangle) \cup \text{peopleIn}_{\Delta}(\langle \text{NS \& typical-case denial of PAS, non-regrettable PAS} \rangle)$

Defn **People₅** : $\text{FinSet}(\text{People}) := \text{peopleIn}_{\Delta}(\langle \text{NS, difficult-regrettable PAS} \rangle) \cup \text{peopleIn}_{\Delta}(\langle \text{NS \& worst-case denial of PAS, non-regrettable PAS} \rangle)$

Assertion 1: $\Delta \text{Of}(\text{peopleIn}_{\Delta}(\langle \text{NS \& typical-case denial of PAS, non-regrettable PAS} \rangle), \vec{F}_{\text{typical}}) >_{\Delta} 0 \boxminus$

From $\langle \text{NS \& typical-case denial of PAS} \rangle$ to $\langle \text{non-regrettable PAS} \rangle$ is a *good change*. Some self-reported religious people, who are not in the intended audience of this argument, will reject this assumption.

Assertion 2: $\Delta \text{Of}(\text{peopleIn}_{\Delta}(\langle \text{NS \& worst-case denial of PAS, non-regrettable PAS} \rangle), \vec{F}_{\text{typical}}) >_{\Delta} 0 \square$

From $\langle \text{NS \& worst-case denial of PAS} \rangle$ to $\langle \text{non-regrettable PAS} \rangle$ is a *good change*. Some religious people, who are not in the intended audience of this argument, will reject this assumption.

Simplifying Assumption 1: $\text{peopleIn}_{\text{IFR}}^{1,*}(\langle \text{NS \& desire for PAS} \rangle) = \text{peopleIn}_{\text{IFR}}^{1,*}(\langle \text{NS \& typical-case denial of PAS} \rangle) \cup \text{peopleIn}_{\text{IFR}}^{1,*}(\langle \text{NS \& worst-case denial of PAS} \rangle) \square$

From the proposition accompanying the introduction of $\langle \text{NS \& desire for PAS} \rangle$, this is equivalent to the assumption that there are no instances of $\langle \text{NS \& best-case denial of PAS} \rangle$ in SQF. This assumption is disputable. It would be better to include the hypothetical set of people who satisfy $\langle \text{NS \& best-case denial of PAS} \rangle$ in Assumption 5 or Assumption 6, which would amount to a strengthening of those assumptions.

$\square \checkmark$ Lemma 70: $\Delta \text{Of}(\langle \text{all people} \rangle, \vec{F}_{\text{typical}}) >_{\Delta} 0$

Defn $\text{People}_{1-5} : \text{FinSet}(\text{People}) := (((\text{People}_1 \cup \text{People}_2) \cup \text{People}_3) \cup \text{People}_4) \cup \text{People}_5$

$\square \checkmark$ Lemma 71: $\langle \text{all people} \rangle = \text{People}_{1-5}$

\square Lemma 4: $\langle \text{all people} \rangle \subseteq ((\text{peopleIn}_{\Delta}(\langle \text{S,S} \rangle) \cup \text{peopleIn}_{\Delta}(\langle \text{S,NS} \rangle)) \cup \text{peopleIn}_{\Delta}(\langle \text{NS,S} \rangle) \cup \text{peopleIn}_{\Delta}(\langle \text{NS,NS} \rangle))$

The four sets on the right hand side are a partition of $\langle \text{all people} \rangle$ since $\langle \text{NS} \rangle$ is the complement of $\langle \text{S} \rangle$.

\square Lemma 5: $\text{peopleIn}_{\Delta}(\langle \text{S,S} \rangle) \subseteq \text{People}_{1-5}$

Follows easily from basic set reasoning, Claim 22, and the definition of People_{1-5} .

Claim 22: $\text{peopleIn}_{\Delta}(\langle \text{S,S} \rangle) \subseteq \text{People}_2$

\square Lemma 6: $\text{peopleIn}_{\Delta}(\langle \text{S,NS} \rangle) \subseteq \text{People}_{1-5}$

Follows easily from basic set reasoning, Claim 23, and the definition of People_{1-5} .

Claim 23: $\text{peopleIn}_{\Delta}(\langle \text{S,NS} \rangle) \subseteq \text{People}_1 \cup \text{People}_2$

Claim 24: $\text{peopleIn}_{\Delta}(\langle \text{NS,S} \rangle) \subseteq \text{People}_{1-5}$

\square Lemma 7: $\text{peopleIn}_{\Delta}(\langle \text{NS,NS} \rangle) \subseteq \text{People}_{1-5}$

Follows easily from basic set reasoning, Claim 25, and the definition of People_{1-5} .

Claim 25: $\text{peopleIn}_{\Delta}(\langle \text{NS,NS} \rangle) \subseteq (\text{People}_1 \cup \text{People}_2) \cup \text{People}_4$

\checkmark Lemma 72: $\text{People}_{1-5} \subseteq \langle \text{all people} \rangle$

✓Lemma 73: $\langle \text{all people} \rangle \subseteq \text{People}_{1-5}$

⊖ ✓Lemma 74: $\Delta \text{Of}(\text{People}_{1-5}, \vec{F}_{\text{typical}}) >_{\Delta} 0$

✓Lemma 75: $\Delta \text{Of}(\text{People}_5, \vec{F}_{\text{typical}}) >_{\Delta} 0$

⊖ ✓Lemma 76: $\Delta \text{Of}(((\text{People}_1 \cup \text{People}_2) \cup \text{People}_3) \cup \text{People}_4, \vec{F}_{\text{typical}}) \approx_{\Delta} 0$

⊖ ✓Lemma 77: $\Delta \text{Of}(\text{People}_4, \vec{F}_{\text{typical}}) \approx_{\Delta} 0$

Claim 26: $\text{Disjoint}(\text{peopleIn}_{\Delta}(\langle \text{NS}, \text{NS \& same-or-better palliative care \& exacerbated guilt} \rangle), \text{peopleIn}_{\Delta}(\langle \text{NS \& typical-case denial of PAS, non-regrettable PAS} \rangle))$

⊕ ✓Lemma 78: $\Delta \text{Of}((\text{People}_1 \cup \text{People}_2) \cup \text{People}_3, \vec{F}_{\text{typical}}) \approx 0$

Chapter 9

Ongoing work

9.1 Web system for collaborative authoring and criticizing of interpreted formal proofs, and a minimal dialogue system

Initially I thought that a sophisticated dialogue system, with rules designed to ensure progress under certain assumptions, would be essential to move forward with this project. With more experience writing interpreted formal proofs, however, it became clear that reasoning faithfully about complicated inelegant structures was already so onerous a task that it would be asking too much of authors to require the extra work of demonstrating that their argumentative moves make progress. This has led me to shift to a relatively simple and lax model of interaction. The end of this chapter contains some notes about the issue of progress.

9.1.1 Related work from Informal Logic

Carneades:

Carneades is a web application in active development “which provides software tools based on a common computational model of argument graphs useful for policy deliberations and claims processing.”[Gor13] It is the application, of those I am aware of, that is most related to the one I am working on, although its focus on propositional defeasible reasoning makes it still only weakly related. In more detail, the principle developer Thomas F. Gordon describes Carneades as a collaborative, online system for (quoting from [Gor13]):

- modeling legal norms and argumentation schemes
- (re)constructing arguments in an argument graph

- visualizing, browsing and navigating argument graphs
- critically evaluating arguments
- forming opinions, participating in polls and ranking stakeholders by degrees of agreement
- obtaining clear explanations, using argument graphs, of the differential effects of alternative policies or legal theories in particular cases

So Carneades is a software system with very broad intended applications, but it is not misleading to say, more concisely, that it is a tool suite for supporting the practice of deliberate defeasible argumentation.

I tried out Carneades. As of 15 Aug 2014, the program comes with only one example, called “Copyright in the Knowledge Economy”. The instructions for the “guided tour” of the example (described as an “opinion formation and polling tool”) certainly have a lot in common with the goals of this thesis:

It guides you step by step through the arguments on all sides of a complex policy debate, providing you with an overview of the issues, positions and arguments in a systematic way. The tool can help you to form your own opinion, if you don't yet have one, or to critically evaluate and reconsider your preexisting opinion, if you do. The tool also enables you to compare your responses with published positions of some stakeholders, such as the official positions of political parties. This can help you to find persons and organizations which best represent or share your views and interests.

However, examining the argument itself one finds that, aside from the tree structure, it is not a great departure from typical natural language arguments. Here is a prototypical example, where “exceptions” means copyright exceptions:

Q4. Should certain categories of exceptions be made mandatory to ensure more legal certainty and better protection of beneficiaries of exceptions?

pro Argument #1: The permitted exceptions should be harmonised so that they are available in all Member States.

- pro Argument #1:
 - Performing the action of harmonizing the exceptions and giving precedence to community law over contracts would achieve a state in which it easier for researchers and students to work in more than one Member State.

- Harmonizing the copyright exceptions would make it easier for researchers and students to work in more than one Member State.
 - Achieving the goal of making it easier for researchers and students to work in more than one Member State would promote the values of efficiency, legal certainty, scientific research and education.
 - In the circumstances: Researchers and students increasingly work in more than one Member State. The patchy availability of exceptions makes their work difficult, because what is lawful in one country is probably unlawful in another. The situation is made worse by the provision of most Member States that contracts, governing the use of digital material, automatically overrides statute law.
- con Argument #2:
 - It is essential that the basic principle of freedom of contract be recognized and preserved by any copyright legislation.
 - Harmonizing copyright exceptions would impair the freedom of contract.
 - Impairing the freedom of contract would demote the values of innovation and the dissemination of knowledge and information.
 - Currently, the lack of harmonization of copyright exceptions facilitates the freedom of contract.

In the argument graphs approach they take, formal logic is not imposed on arguments. Instead, the dialogue features, together with the fundamental notions of an argument attacking or defending a proposition, must be used by one arguer to try to make the other's reasoning seem less sound.

9.1.2 Design of a web system

This section describes work in progress.

Interpreted formal proofs are written using a web-based IDE (integrated development environment), where the document is tree-structured except for some leaves that contain natural language text, which is scanned for symbol ids to insert references. There is an auto-complete feature for already-declared symbols. Declarations (axiom, lemma, or new symbol introduction) can be tagged, to make groups of declarations, in order to more-concisely specify a subset of declarations that should be used to prove a lemma¹.

¹As I mentioned earlier in Section 2.4, this has so-far been necessary when there are a large number of declarations, due to the non-goal-directed nature of the saturation-based first order theorem provers that I have used.

An author of an interpreted formal proof may grant edit privileges to other users, and simultaneous multi-collaborator editing, versioning, and unbounded undo will be implemented with the help of a *realtime framework*, such as the Google Drive Realtime API. The vision is that interested people who don't know formal logic will be able to contribute to interpreted formal proofs by writing and improving the many required sections of natural language text, both for the language interpretation guide entries and for introductions to arguments.

When a user begins to criticize an author's interpreted formal proof, a dialogue data structure is created. It stores, first of all, the critic's current critique of each axiom, which includes one of the following stances:

- accept - the critic commits to having only personal interpretations of the current language that satisfy the axiom.
- weakly reject - the critic commits to having some personal interpretations of the current language that satisfy the axiom, and some that falsify it.
- strongly reject - the critic commits to having only personal interpretations of the current language that falsify the axiom.
- semantics criticism - the critic submits at least one symbol used in the axiom whose language interpretation guide entry is too vague for them to evaluate the axiom - that is, to take one of the previous three stances.

A critique of any of the latter three categories may have attached to it declarations that are owned by the critic (see Section 2.2.1 for details). The dialogue data structure also stores responses to those critiques by the author of the interpreted formal proof, which may also include new declarations. The author may make changes to the original declarations of their interpreted formal proof that are local only to one dialogue.

The greatest foreseen challenge is when the author wishes to make a change to their interpreted formal proof that affects all ongoing dialogues² (hereafter: a change to their interpreted formal proof's *root document*), especially for changes that are not a direct response to a criticism. That includes improvements initiated by the author to the wordings of language interpretation guide entries (which will happen very often), as well as major structural changes to the proof to fix an earlier-made poor formalization decision (i.e. refactoring, as it's called in software engineering). Such changes can subtly affect the meaning of critiques, or in the worst case render current dialogues unintelligible. My current best idea to address this is as follows. When an author wishes to make a change to a root document that is currently involved in at least one dialogue, they are asked to

²Note that dialogues will happen slowly, since the work is being done in users' free time.

make a claim about the change’s effect on each ongoing dialogue, in particular whether they expect the change to require certain specific kinds of adaptations by the critics of the current state of their criticism. Fortunately, it should not ever be necessary to make adaptations to the entire dialogue, since the author’s change to the root document can be recorded in each dialogue and viewed in the same way as a change that was initiated as a response to a critique.

9.2 Obstacles for this project

The true reason for this straying from the portal of knowledge³ 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])

I have come to believe that there is only one significant obstacle against this project gaining interest, which is the difficulty of writing and reading interpreted formal proofs. The proofs in this paper are tedious to read, even in HTML, and they were tedious to write as well. I do not have a perfect solution in mind for this problem. My current approach is to eliminate barriers to entry for use of the software system (e.g. programming experience shouldn’t be necessary to author or criticize an argument, and one should be able to get started immediately, on any operating system, without installation), and to minimize editing friction as much as possible (e.g. type-aware autosuggest, as found in modern IDEs for strongly-typed programming languages).

For a long time I have been concerned about the practical effect of authors and critics who do not argue “in good faith”, according to the ideals set out in Section 2.2. I no longer believe that uncooperative *authors* will be a significant concern. Due to the fundamental difficulty of writing an argument as a computer-readable formal deductive proof (even with an excellent user interface), I expect to have interest only from authors who will want to write their arguments so that they are as strong as possible according to the critics whose reasoning they respect the most. In contrast to authoring a new interpreted formal proof, writing a simple criticism of an existing proof (Section 2.2.1) is by design not difficult, so there is greater potential for frivolous criticisms and other time-wasting uncooperative behavior. However, formal logic provides a natural notion of progress by which most earnest criticisms will “make progress”, namely the proof-theoretic strengthening of a system of axioms, and the growing proof-theoretic independence relations that result from

³Leibniz is referring to his imagined *characteristica universalis*, or any similar project.

a critic weakly-rejecting an axiom, or an author weakly-rejecting an axiom proposed by a critic (recall this means a commitment to the independence of the axiom, which has consequences of independence of other sentences). Of course, when an author's intended interpretations of their language are infinite structures, this notion of progress is not necessarily terminating, even when the language is never non-conservatively extended, since for non-propositional first-order languages \mathcal{L} , there are infinite families of \mathcal{L} -theories $\Gamma_1 \subset \Gamma_2 \subset \dots$ such that Γ_{i+1} is strictly stronger than Γ_i for all i . I do not expect that technical possibility to arise in practice *except* when it is accompanied by non-conservative extensions of the language. There is a correspondence between the non-conservative extensions case and the familiar experience in informal argumentation where a dialogue is never ending due to the *scope* of the argument being repeatedly expanded. The theoretical issue is that it may be difficult to distinguish by any practical technical test between those non-conservative language extensions that are necessary to state a criticism, and those that are merely stalling. Whether this occurs often in practice, and if so how it effects the goals of the project, remains to be seen.

Bibliography

- [AFKP13] Itai Ashlagi, Felix Fischer, Ian A Kash, and Ariel D Procaccia. Mix and match: A strategyproof mechanism for multi-hospital kidney exchange. Games and Economic Behavior, 2013.
- [BCD⁺11] Clark Barrett, Christopher L Conway, Morgan Deters, Liana Hadarean, Dejan Jovanović, Tim King, Andrew Reynolds, and Cesare Tinelli. CVC4. In Computer aided verification, pages 171–177. Springer, 2011.
- [BdRV02] P. Blackburn, M. de Rijke, and Y. Venema. Modal Logic. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, 2002.
- [BHO75] P. J. Bickel, E. A. Hammel, and J. W. O’Connell. Sex bias in graduate admissions: Data from Berkeley. Science, 187(4175):398–404, 1975.
- [BS13] Berit Brogaard and Joe Salerno. Fitch’s paradox of knowability. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Winter 2013 edition, 2013.
- [CC14] Massimiliano Carrara and Daniele Chiffi. The knowability paradox in the light of a logic for pragmatics. In Recent Trends in Philosophical Logic, pages 31–46. Springer, 2014.
- [DF99] Peter Donnelly and Richard D Friedman. DNA database searches and the legal consumption of scientific evidence. Michigan Law Review, pages 931–984, 1999.
- [DH50] Richard Doll and AB Hill. Smoking and carcinoma of the lung (reprint). Bulletin of the World Health Organization, 77(1):84–93, 1999 (original 1950).
- [Don05] Peter Donnelly. Appealing statistics. Significance, 2(1):46–48, 2005.
- [Efr05] Bradley Efron. Bayesians, frequentists, and scientists. Journal of the American Statistical Association, 100(469):1–5, 2005.

- [F⁺11] Stephen E Fienberg et al. Bayesian models and methods in public policy and government settings. Statistical Science, 26(2):212–226, 2011.
- [Far93] William M. Farmer. A simple type theory with partial functions and subtypes. Annals of Pure and Applied Logic, 64(3):211–240, November 1993.
- [Gor88] Thomas F. Gordon. The importance of nonmonotonicity for legal reasoning. In F. Haft H. Fiedler and R. Traunmüller, editors, Expert systems in law: impacts on legal theory and computer law, Neue Methoden im Recht. Attempto Verlag, 1988.
- [Gor13] Thomas F Gordon. Introducing the Carneades web application. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, pages 243–244. ACM, 2013.
- [Kad08] Joseph B. Kadane. Statistics in the Law : A Practitioner’s Guide, Cases, and Materials. Oxford University Press, USA, 2008.
- [KV13] Laura Kovács and Andrei Voronkov. First-order theorem proving and Vampire. In Computer Aided Verification, pages 1–35. Springer, 2013.
- [Kva95] Jonathan Kvanvig. The knowability paradox and the prospects for anti-realism. Noûs, pages 481–500, 1995.
- [LCMJ10] M. Lynch, S.A. Cole, R. McNally, and K. Jordan. Truth Machine: The Contentious History of DNA Fingerprinting. University of Chicago Press, 2010.
- [LCS11] G.W. Leibniz, S. Charlotte, and L. Strickland. Leibniz 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. Philosophical Papers and Letters. Number v. 1 in Synthese Historical Library. D. Reidel Publishing Company, 1976.
- [Man96] María Manzano. Extensions of First Order Logic. Cambridge University Press, 1996.
- [Pea09] Judea Pearl. Causality: 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. Argument & Computation, 1(2):93–124, 2010.
- [Res85] N. Rescher. The Heritage of Logical Positivism. CPS publications in philosophy of science. University Press of America, 1985.
- [SE05] George Davey Smith and Matthias Egger. The first reports on smoking and lung cancer: why are they consistently ignored? Bulletin of the World Health Organization, 83(10):799–800, 2005.
- [Ses07] Neven Sesardic. Sudden infant death or murder? a royal confusion about probabilities. The British Journal for the Philosophy of Science, 58(2):299–329, 2007.
- [Sor13] Roy Sorensen. Vagueness. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Winter 2013 edition, 2013.
- [SPMS09] Yun S Song, Anand Patil, Erin E Murphy, and Montgomery Slatkin. Average probability that a “cold hit” in a DNA database search results in an erroneous attribution. Journal of forensic sciences, 54(1):22–27, 2009.
- [Sut09] G. Sutcliffe. The TPTP Problem Library and Associated Infrastructure: The FOF and CNF Parts, v3.5.0. Journal of Automated Reasoning, 43(4):337–362, 2009.
- [SWC00] Mark E Stickel, Richard J Waldinger, and Vinay K Chaudhri. A guide to SNARK. Technical report, SRI International, 2000.
- [Tal13] William Talbott. Bayesian epistemology. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Fall 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. Bulletin of the World Health Organization, 83(2):144–145, 2005.
- [Ven98] Yde Venema. Review of “Extensions of first order logic”. The Journal of Symbolic Logic, 63(3):pp. 1194–1196, 1998.
- [Wal08] D.N. Walton. Informal Logic: A Pragmatic Approach. Cambridge University Press, 2008.

- [Wal11] D.N. Walton. Finding the logic in argumentation. International Colloquium, Inside Arguments, 2011.
- [WG85] Ernest L Wynder and Evarts A Graham. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma: A study of six hundred and eighty-four proved cases. Jama, 253(20):2986–2994, 1985.
- [WK95] D. Walton and E. Krabbe. Commitment in Dialogue: Basic concepts of interpersonal reasoning. State University of New York Press, Albany NY, 1995.