Computing moral hypotheticals

by

Dylan Alexander Holmes

B.S., Wichita State University (2012)
M.S., Massachussets Institute of Technology (2017)

Submitted to the
Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Electrical Engineering and Computer Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2021

Signature of Author .....
Department of Electrical Engineering and Computer Science
August 27, 2021

Certified by .....
Randall Davis
Professor of Computer Science and Electrical Engineering
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Computing moral hypotheticals

Dylan Alexander Holmes

Abstract

Our moral judgments depend on our ability to imagine what else might have happened: we forgive harms that prevent greater harms, we excuse bad outcomes when all others seem worse, and we condemn inaction when good actions are within reach. To explain how we do this, I built a computational model that reads and evaluates short textual stories, computing hypotheticals in order to make moral judgments.

I identify what specialized knowledge we need in order to know *which* hypothetical alternatives to consider. I show how to connect abstract knowledge about moral harms to the particular details in a story. Finally, I show how the system can assess outcomes in a purely qualitative, human-like way by decomposing outcomes into their harmful components; I argue that—as in real life—many outcomes are incomparable.

I support my theoretical claims with references to the cognitive science and philosophical literature, and I demonstrate the system's explanatory breadth with diverse examples including *escalating revenge*, *slap-on-the-wrist*, *preventive harm*, *self-defense*, and *counterfactual dilemma resolution*.

The key insight is that hypothetical context modulates understanding. With this system, I shed light on what is needed to grasp hypothetical context as effortlessly and automatically as we humans do. And I lay the groundwork for moral reasoning systems that are as nuanced, imaginative, and articulate as we humans are.

Thesis supervisor: Randall Davis
Title: Professor of Computer Science and Electrical Engineering

Acknowledgements

So many people helped me throughout the course of this work that I can only begin to acknowledge the appreciation that I owe and feel.

I am grateful for the advice and critiques of my committee members, Randall Davis, Gerald Sussman, and Peter Szolovits. Randall Davis was a steadying influence and indefatiguable source of support and encouragement from the moment we met. He helped me keep an even keel through several downturns during the course of this thesis, and—with a keen ear for crisp phrasing and a sense of how arguments can be misconstrued—he helped me find just the right ideas and the words to go with them.

Peter Szolovits helped shape this work long before he became an official mentor. Every time I gave a presentation with Peter in the audience, he invariably asked a good question that led me to improve some part of my system. This happened so consistently that eventually the light bulb went off and I asked if he would join my committee. Peter agreed, and the rest is history. I am especially grateful for his insightful comments on the knowledge engineering aspects of my system; through his urging and encouragement, I was able to strengthen my arguments on scalability and robustness.

Gerald Sussman showed me that play is essential to understanding. It was through our conversations—conducted always over tea, and often on a spur-of-the-moment visit—that I began to think of this thesis as a way to seriously play with big ideas about moral reasoning and imagination. Also, once I had finished the first draft, Gerald was the first person to understand what I had accomplished, what it was really about[1] and what the new and important contributions were.[2] I made sure to take notes when he explained it to me.

Finally, I feel tremendous gratitute toward Patrick Winston, my advisor when this whole work began. I've had the experience of a lifetime at MIT, all of which leads back to Patrick: I met him almost as soon as I moved to Cambridge, MA in 2012, trying to go from zero to AI researcher the way some people move to

[1] Using computational concepts to precisely articulate moral concepts.

[2] For example, demonstrating a computational moral theory which compares harms qualitatively instead of reducing everything to numerical scores.

Hollywood to make it in movies. I remember how intimidated I felt showing up at his office door for advice. At the time, he pointed out that since I wasn't in school at the time, let alone an MIT student, he probably couldn't help me. He then added, offhandedly, that he was terribly short staffed for his AI course starting in a few weeks, and could I come back tomorrow on an exceptionally provisional basis? Of course I accepted, and I've been part of his team ever since.

Patrick was kind, but not showy. He had an uncanny talent for distilling the exact essence of any presentation, algorithm, or half-baked suggestion in a few words[3]. And Patrick believed that a university could absolutely be the place where you fulfill a dream or spark a lifelong interest; where you learn about the world and how to get along; where you take on and solve hard, real-world problems. Of course he did—that's what his whole life at MIT was about, and that's what he put into practice for so many generations of students. In so many ways, he was visionary, who put in the work every day to make it happen. He was a leader of unmatched integrity, courage, and optimism, and it was such an honor to work with him. I carry his legacy with me today.

I am grateful to my family, that far-flung constellation of people who I return home to. I am in particular grateful to my parents, who instilled in me a resillient sense of humor, a healthy disrespect for the proper way of doing things, and the benevolent sense that they could care less about my path in life provided I was happy about it. May all children be so lucky. I am, of course, also grateful to my brother Cameron, whose mischievous sense of humor, gentleness, and clear-eyed sense of justice inspire me to this day. His memory lights my way forward.

I am grateful to many colleagues who have inspired, supported, and challenged me over the years. I'd like to especially acknowledge Doug Riecken for his tireless support of foundational AI research and for championing my work in particular. I've been lucky to know him as a fellow traveler and friend.

[3]This is surprisingly hard to do without flattening the idea or missing what's new about it.

Pat Langley helped sharpen my approach to cognitive systems and helped me find my way in the wider AI community. With Pat's advice—conveyed with his characteristically arch sense of humor—I was first able to make my work intelligible to researchers in neighboring fields, which helped strengthen it considerably.

Aaron Sloman inspired me to think beyond sense data, to be alert to how much we grasp about affordance and impossibility. He has been a gracious internet pen pal over the years, and to the extent that I have been able to make any philosophical headway in artificial intelligence, it is due to his analytical influence.

I'd also like to express my thanks to my colleagues Maria Rebello, Howie Shrobe, Jamie Macbeth, Josh Tenenbaum, Laura Schulz, Danny Weitzner, Henry Lieberman, Edwina Rissland, Rogelio Cardona-Rivera, Kimberle Koile, and Bob Berwick for supporting and inspiring me in ways both large and small throughout my time at MIT.

To my wonderful friends—what an adventure it's been. I'd especially like to acknowledge Adam Kraft, who was in the trenches with me for the many years we shared a grad student office. In Adam, I found a real mensch and a friend for life; our late night conversations on science, society, and the future were among the best parts of my MIT experience. They have shaped who I am, and I know they will continue to shape what comes next.

Thank you to Jessica Noss for—so much. Thank you for all the puzzles, the adventures, the stories, and the songs; for keeping the 6.034 ethos when everything was on fire, for setting the gold standard as head TA and factotum for life, and for all the fun we've had since.

Thank you to Leilani Gilpin for being such a steadfast friend and a powerhouse of an academic peer. Whether we were cheering on each other's presentations, reassembling our self-worth after a gruelling class, or coordinating an AI symposium, Leilani showed up completely. I especially admired her commitment to fostering constructive critiques and big, transformative projects in academia; tomorrow's leaders will flourish if they take notes from her.

Thanks to Kelsey Allen, science friend extraordinaire, for many hangout games ranging from bananagrams to baking, and for discovering the unusually helpful co-working strategy of meeting up to code our separate projects side-by-side in absolute silence. I don't know how I would've gotten it all done otherwise—somehow, it really helps to have a good friend nearby. When we met at the summer school for the Center for Brains, Minds, and Machines, she, along with Leo Casarsa de Azevedo, Nicole Rafidi, and I, spent a wonderful couple weeks cramming cross-disciplinary knowledge as much as possible by day and gamboling around looking at bioluminescent sea critters by night. I know of no better recipe for fast friends, and I am so lucky to know them all.

Thank you to Cynthia Solomon, Margaret Minsky, Gloria Rudisch Minsky, and the rest of the Minsky crew for welcoming me into the fold. My memories of our time together glitter with little moments—our Golden Gallery AI demo, the singing valentine, playing with LOGO lights, listening to Marvin's improvisations. I am grateful in particular to the late Marvin Minsky, who spontaneously invited me to visit his house in Massachussetts when I—as a young student—breathlessly cold called him from my house in Kansas. I accepted the invitation to visit (the first of many, as it turned out) and I soon learned that he was always like that—spontaneous, unexpected, generous, and inventive. To this day, he has shaped how I think and how I see the world.

Thank you to Suri Bandler, who got it exactly right that you can live by your principles even when it's hard, and you can have fun in school even if it's chic to be miserable—-I've enjoyed walking a mile in your boots. Thank you, Avril Kenney, for cooking up muffins and life plans and always taking the time to understand what I mean; thank you Suzannah Fraker, for all the colors and music; thank you Héctor Vázquez Martínez, for showing up time and time again and getting the letter to Garcia; thank you Yida Xin, for asking the important questions at just the right time; thank you Caroline Aronoff, for all the collaborative stories; thanks Çağrı Zaman, for helping me see like you do; thank you, Judy Yates, for your unwavering thoughtfulness and for welcoming me in; thanks Ben Yuan for your contagious enthusiasm and for graciously answering my dozens of questions; thanks Zach Visco, for always encouraging me and

helping me find my way forward; thanks Ronit Langer, for sharing space for reflection, action, and inspiration when times were changing; thanks to Clarence the rat for being so gentle with your teeth; thanks Sila Sayan for your humor, your wise questions, and your sunshine; thanks so much, Jing Fan, for coming by to say hello; thank you Terri Hinton, for welcoming me to the tribe; thank you Will Rogers for showing me other ways of knowing; thank you Zhutian Yang, Patrick's lieutenant, who showed how much is possible when inspiration strikes, and who took the Genesis tribe to another level.

Most of all, I am grateful to Robert McIntyre, who supported me in every way on this project and who inspires me every day with his compassion, ingenuity, and thoughtfulness. I am honored to be on this adventure with him.

The great instrument of moral
good is the imagination.

—PERCY SHELLEY

Contents

1 Introduction

1.1 An Iceberg Theory about Thinking

What makes *Romeo and Juliet* a tragedy? It's not just the parade of disasters. In part, it's that things could have easily turned out differently: Consider Juliet, who plans to fake her own death and sends Romeo a note warning him. When plague breaks out, the message-bearer is quarantined and doesn't deliver the note in time. Romeo believes Juliet to be truly dead. If it weren't for a number of bad turns like that, things might have turned out all right in the end.

These poignant alternatives enrich the story. And somehow we all instinctively sense them, side paths laid out alongside the main road of the story. How do we do that? Or, to ask questions like an engineer, what are the processes involved? How do they interact? How do we know to consider some alternatives and not others? What knowledge (about mail carriers, plagues, and false beliefs) do these processes need, how is that knowledge represented, and how is it deployed?

These questions are crucial, and not just because they can explain how we enjoy satisfying stories. In fact, these questions about alternative stories expose some of the fundamental machinery of human cognition.

Our understanding of *Romeo and Juliet* is built out of much more material than what we explicitly see. We arrive at poignancy in part because of alternatives that we perceive are imminently within reach. These alternatives weren't in the story; we had to build them to see them. In other stories, we react with suspense—at an imagined outcome—or surprise—when the story defies the odds or our expectations. And we judge characters not only based on what they do, but also on the many things they *could've* done instead. In short, *we understand stories because we are intimately and expertly aware of available alternatives.*

This is an iceberg theory about thinking. It is the idea that although we *start* by digesting what's explicitly in front of us, this is only the most superficial part of the analysis. Below the surface, we have many processes for extrapolating

outcomes, filling in implicit connections, noting possibilities, and imagining alternatives. The *hypotheticals* we construct supply more grist for understanding than the facts at hand do. If you look for this sort of pattern, you begin to see it everywhere, not just in story understanding. Vision is like this, for example. The psychologist James Gibson[4] points out that seeing is purposeful—not for turning the flat retinal image into a solid model of the world, but rather for quickly deciding what's there, what it's like, what's possible, and what to do. Under deadly evolutionary constraints, each species hones a specialized perceptual apparatus tuned to certain *affordances* in the environment—aspects of space and possibilities for action that matter to that animal. Rather than simply digesting an image of the world into a solid model, practical vision involves aligning sensory information with latent possibilities for what could be next.

Conversation fits this pattern, too. Understanding everyday conversation requires *implicatures*—you can't understand what someone is saying without strategically understanding what they *could've* said instead. Consider condemnation by faint praise: "How was the play?" "About three hours long.". If you understand only the explicit meaning of everything said in a conversation, you miss most of what people have to say.

In my view, engineers and professionals of all stripes ought to be curious about hypothetical reasoning. What-if scenarios appear as a cornerstone of law ("What if the assailant had brandished a knife?"), medicine ("What if the patient's T-cell count decreases?"), and business ("What if we invest here?"). Imagine if we could build machines that could imagine hypotheticals as effortlessly and cogently as we can.

Teachers, students, and knowledge engineers ought to be curious, too. In any domain, deep understanding is in part the ability to understand which differences make a real difference. And what better way to uncover latent knowledge than to ask a few well-chosen what-if questions? Imagine a system that can learn and self-correct by asking the right what-if questions. I think we could build more robust knowledge systems, and teach and learn more effectively, by organizing knowledge in terms of such what-ifs. Patrick Winston's (Winston, 1970) idea of

[4] Gibson (2014)

near-miss learning is an important step in that direction.

I am especially curious about the computational processes involved in hypothetical reasoning. On the face of it, we seem capable of feats of outright computational wizardry. Just think: even when we are very young, we effortlessly react to stories with feelings of suspense, surprise, poignancy, etc.—all feelings based on a sense of what could happen. But how do we know *which* hypotheticals to think of? Presumably we don't constantly generate all possible spinoffs and filter for the ones that are interesting—that's too expensive—so we must have some regulatory processes or reflexes that help us decide when to fire up our imagination. How do they work? How can we know when it's an apt time to imagine, without imagining first?[5] And another question: once we've decided to imagine, how do we fill in the details? Because we know so many things about the world, how do we avoid getting bogged down imagining details that fit but are useless? Evidently, we are able to zero in on just the details we need, on an appropriate level of abstraction. Consider, for example, the blocks in Figure 1.1. You can even imagine how to grasp and manipulate the blocks, despite the fact that the figure is physically impossible. Evidently you are capable of faithfully representing *aspects* of the world without necessarily *simulating* the whole thing in rigorous detail. How do we know which aspects to include in our imagined scenarios, and how do we fill in additional details as needed?

And now, asking questions about imagination, we begin to encounter some of the deepest questions about how our knowledge is organized and re-organized over the course of a lifetime. Every adult develops an astonishing amount of intuitive, utterly mundane knowledge of the world—like the fact that you can pull things using a piece of string, but not push them (McCarthy et al., 1960; Spelke and Kinzler, 2007). Somehow, we manage to deploy this knowledge to solve problems in the world: What can I use to plug this leak? What should I call my new restaurant? How do candy canes get their stripes[6]? Given that we can imagine hypothetical alternatives on the fly, to varying levels of detail, we must be able to marshall that knowledge, to align concepts analogically, to

[5] See Minsky (1994)'s discussion of negative knowledge and its role in directing thought.

[6] See Magid et al. (2015).

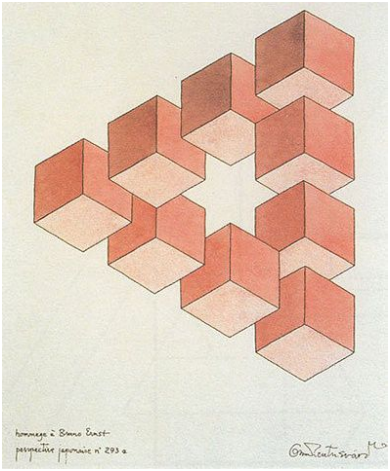


Figure 1.1: We can envision possibilities at an appropriate level of abstraction. Here, a physically-impossible triangle provides many affordances for 3D spatial grasping and manipulation, such as swapping adjacent cubes. The physical impossibility is no obstacle because we have a representation that abstracts over precise physical details like absolute coordinates.

re-represent knowledge in these *ad hoc* ways, with tremendous speed, flexibility, and accuracy. In other words, we are somehow able to organize our knowledge auspiciously to find good answers and generally avoid bad ones; note that it is once again not enough to simply search over the sum of our knowledge as represented in canonical form in some universal language—there is simply too much knowledge, and too much detail within it, for that to work on an everyday basis. The good ideas are just too thinly represented among all possible expressions. On the other hand, how might we pluck out just the right knowledge we need, isolating just the right features for selection, without that kind of exhaustive census? And what are its limitations? I propose some preliminary suggestions in Section 2.1.

1.2 The computational theory

I have developed a computational theory to explain how we see hypothetical context and use it to ground our judgments. In this thesis, I describe what specialized

knowledge we require, particularly knowledge of possibilities and impossibilities, and how we evaluate hypothetical scenarios.

The corresponding computer program[7] applies these hypothetical reasoning principles to make moral arguments. When reading a short text-based story, the system can evaluate actions by referring to outcomes that could have otherwise happened. For example, the system can excuse *preventive harms* (Fig 1.2), producing explanations like "Although it was wrong to swat at a friend, it was excusable because a wasp sting *would have been* worse.", or "Although the intruder didn't cause serious harm, they plausibly *could have*—the counter-attack was self-defense."

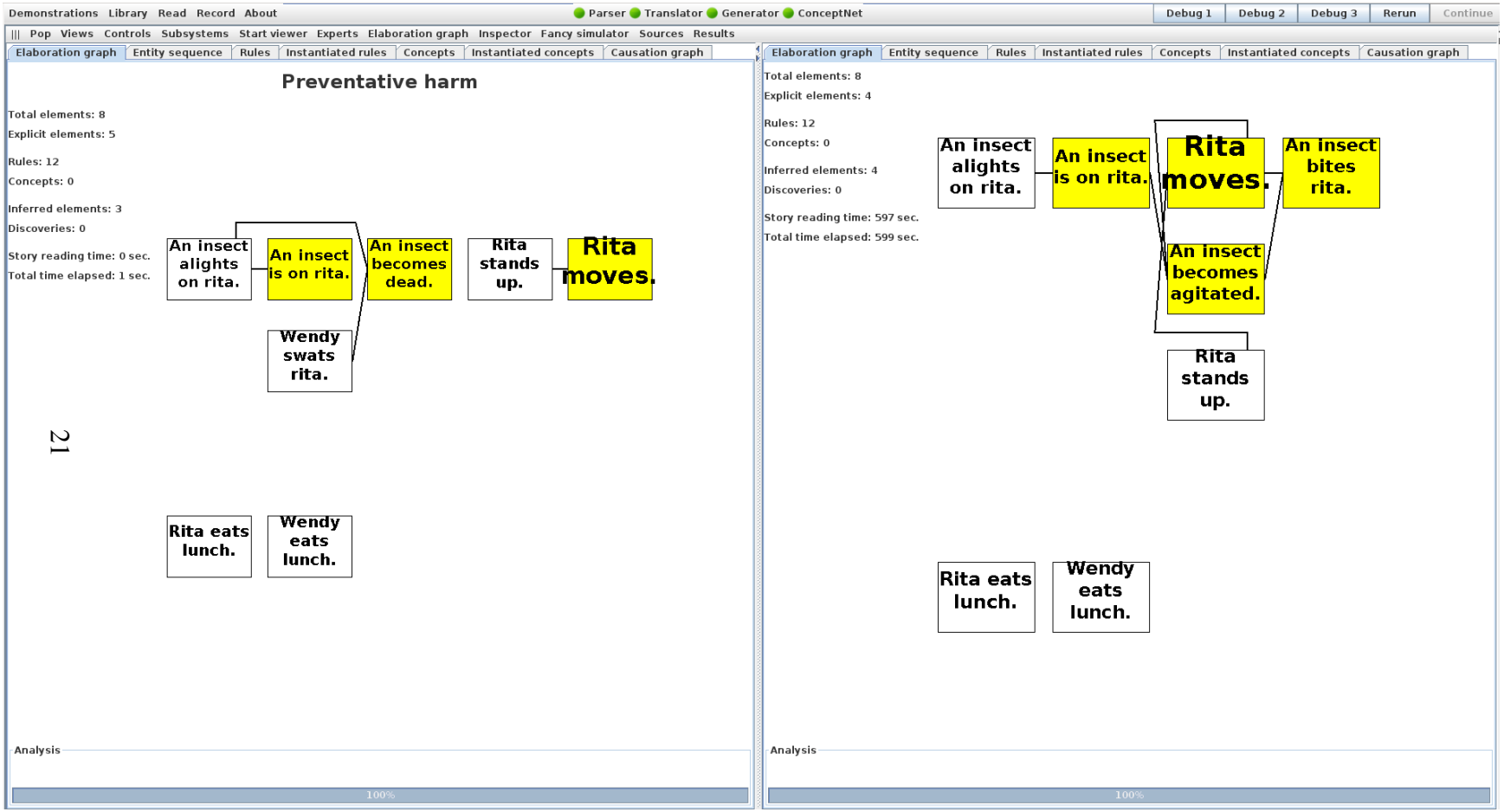
In this way, the system's reasoning is based not only on what happens in the story, but on what *could* have happened. Its domain of knowledge is moral reasoning, which I chose as my area of particular focus because moral problems are interesting and many moral concepts are naturally expressed in terms of hypotheticals (e.g. culpability, self-defense, extenuating circumstances, forebearance). To arrive at its judgments, the system identifies the moral content in a story, makes qualitative comparisons, and performs efficient searches through possible what-if scenarios. Besides the preventive harm examples noted above, the system exhibits other hypothetical reasoning competences. These include recognizing relative degree of harm (in thematic concepts such as *escalating revenge* or its opposite, *slap on the wrist*), and reasoning about what-if features in moral dilemmas (such as "You should probably jump into a river to save a drowning person, even if it would ruin your jacket, even if that jacket were expensive, and especially if the drowning person were a child.").

The point of these computer-generated judgments is not that they are inarguably morally correct, but that they are produced by processes and representations that resemble our own and that are humanly plausible. The exact value system can vary, as it does among human beings. In fact, by using the system's flexible representational scheme, the user can describe value systems of differ-

[7] **A note on implementation** The program described in this thesis was written in Clojure (a variant of LISP that runs on the Java virtual machine) with a few minor plumbing details done in Java. It builds upon our group's research program, the Genesis Story-Understanding System, which is written in Java. I provide a brief overview of the Genesis system in Appendix B.

ent people, cultures, or temperaments and observe the effect on the resulting computational judgments.

I built this system because I believe that hypothetical reasoning is a key part of how we humans understand the world around us. It is at the core of our ability to react to stories with suspense, surprise, and poignancy; our ability to reason about the world; and our ability to imagine, develop new ideas, and learn.



Rita and Wendy eat lunch. An insect alights on Rita. Wendy swats Rita. Rita stands up.

Figure 1.2: The program can use hypothetical reasoning to excuse harmful actions. In this particular scenario, a person swats at their friend to scare away a wasp. The system notes that swatting is harmful, identifies the dire possibility of a wasp sting, notes that the swat was preventive, and declares the harm excusable.

2 Hypothetical alternatives ground our moral judgments

To get a handle on hypothetical reasoning, we'll look at moral reasoning, where it plays a pivotal role. In moral reasoning, we are constantly citing hypotheticals: whether the wrongdoer had any other choice, whether a bystander could have intervened, whether self-defense is justified.[8] These judgments depend not only on what explicitly happens—they depend on what could have been. By studying what knowledge and processes are required to understand moral hypotheticals, we will distill principles about hypothetical reasoning in general.

In this chapter, I present a theory about how hypotheticals ground our moral judgments. The corresponding computer program models this moral reasoning behavior by generating, analyzing, and comparing hypothetical scenarios. The program, built on top of the Genesis story-understanding system, reads text-based stories. When it reads a story, it identifies the harms that occur and compares them against the harms that might have happened otherwise. Through such comparisons, the program can cite hypotheticals to support moral evaluations, e.g. excusing harmful behavior when it prevents a greater harm from occurring, noticing and condemning inaction when helping is easy, and justifying self-defense in light of the harm it prevents.

I use the concrete example of moral reasoning to shed light on our hypothetical reasoning competence as a whole. The result is best expressed as a set of questions and general principles, which I discuss in the sections that follow:

What knowledge do we need? We must know incisively what can and can't happen otherwise.

How do we analyze hypothetical scenarios? We connect the details in the story to general abstract categories, such as bodily harm.

[8] Philosophers, too, have thought that hypotheticals are important to moral reasoning. See (Nozick, 1974, p. 84)'s characterization of blackmail and Jackson (2016) for example.

How do we compare hypothetical scenarios? We compare scenarios qualitatively, comparing features rather than magnitudes.

2.1 We know what could and couldn't happen

What do we need to know in order to reason hypothetically? We understand much more than what's explicitly in front of us. We understand what could have been. We understand these hypothetical alternatives so clearly and precisely that they ground our basic moral judgments and elicit powerful emotional responses—suspense, surprise, poignancy, and the rest. We don't say: 'What I've just read *might be* poignant', or 'Deflecting the knife blow *might be* self defense'. We say: just *look* at what could have happened—as if it's right there for everyone to behold.

This understanding requires *incisive* knowledge of alternative possibilities: the possibilities must be cogent enough to support our judgments. Take the case of self-defense, for example. Here is a story about a quarrel that takes place in a bar:

> George, Alex, and Martha are persons. Martha is George's spouse. Alex is George's lover. Martha and Alex despise each other. Martha encounters Alex and George at a bar. Martha yells at Alex. Alex brandishes a knife. Martha shoots Alex, then confronts George.

Reading the story, we can argue that Martha is acting in self defense when she shoots Alex. The interesting question is not whether this argument is *morally correct*, but what processes we use to construct it. We can appeal to a hypothetical scenario: Martha was in imminent danger—she might've been hurt *if* she hadn't intervened. The hypothetical scenario becomes evidence we use to determine if Martha's violence is justified.

The theory is that in any situation, we have specialized knowledge about what could possibly happen. We use this knowledge to fill in the details of imagined hypothetical scenarios—e.g., what might happen if Martha had *not*

shot Alex—which inform our assessment of what actually occurs. If Martha could've been hurt, her violent action might be excusable as self-defense.

To encode knowledge of possibilities in my moral reasoning system, I developed two types of commonsense inference rule, extending the rule types provided by the base Genesis story understanding system. *Presumption rules* encode what might happen (such as "if you brandish a knife when angry, you might hurt someone"), while *censor rules* encode what can't happen (such as "if a person is unconscious, they cannot harm you").

Thus if we introduce two commonsense rules[9] into our knowledge base

```
If xx brandishes a knife and xx is angry with yy, then
xx could presumably stab yy.
If xx is dead, then xx cannot hurt yy.
```

the system can begin to reason about the alternatives in this story the way we do. We can ask the system to read the story, then prompt it with a what-if question: "What would happen if Martha didn't shoot Alex?"

The system answers the question by first imagining the alternative scenario. It makes the modification in a straightforward way, by deleting the sentence "`Martha shoots Alex`". Then, it rereads the modified story. Because the system has presumptive knowledge to fill in the gap, the modified story is not merely shorter—new things happen. The system infers that if Alex brandishes a knife and Martha does not use a gun in response, then "`Alex presumably stabs Martha`".

Having described how the system works on a particular example, let me describe how it works in general.

My system is built on top of the Genesis story-understanding system, which provides the basic substrate for reading a story, filling in gaps with commonsense

[9]Note that in the Genesis system, rules are all expressed in natural language, and by convention double-letter pairs such as `xx` and `yy` are understood as variables. When the system sees a rule with the keyword *could* or *presumably* ("...`xx could hurt yy`"), it handles it as a presumption rule. Similarly for *cannot* and censor rules. For more details, see Winston and Holmes (2018).

information, and identifying themes (see Appendix B). In order to understand a hypothetical question like "What would happen if Martha didn't shoot Alex?", the system first reads the story. Then the user can prompt with a question of the form "What if this event didn't happen?" The system creates a copy of the story in a new story-context, deletes the event, and analyzes the story anew.[10] The Genesis system's knowledge base fills in commonsense details, with explicit facts and inferences taking precedence over (and overruling) imagined hypothetical details.

Because of the presumption and censor rules, deleting an event from the story does not merely make it shorter; new things can happen. The *presumption rules* fill in new details in a provisional way: they are fragile default assumptions, which can be overwritten by other rules, or elaborated with additional detail. The *censor rules* encode a kind of negative knowledge (Minsky, 1994) about things that cannot happen. They control the details of the imagined scenario by precluding certain other rules from firing. The interplay of 'positive' and 'negative' knowledge enables the system to anticipate plausible alternative scenarios[11].

Possibilities as presumptive inference Presumption rules and censor rules capture knowledge about a particular kind of possibility—possibilities that are so imminent, we automatically infer them. This kind of possibility is important for understanding our automatic grasp of hypothetical context, so this is the kind that I model here.

Of course, in other applications of hypothetical reasoning, possibilities may be significantly more complex, requiring more than just inference: How would Shakespeare's *Macbeth* turn out if Lady Macbeth weren't greedy? Answering the Lady Macbeth question is less like filling in commonsense information and

[10] See Appendix B for implementation details.

[11] See Holmes (2017) for more details about presumption rules and Winston and Holmes (2018) for more details about censor rules. Also, ibid., you can see an example of the interplay between positive and negative knowledge. In a model of 'hyperpresumption' in schizophrenia, we speculate that *under-regulation* by negative knowledge causes atypical inferences (presumptions that aren't properly suppressed by contextual cues) and overriding presumptions (presumptions that supplant available information.)

more like constructing an answer anew; it requires more invention, deliberative imagination, and a greater variety of knowledge.

The rules in my system encode the simpler kinds of inferences. They are, by design, one-step inferences such as "If you fall into water, you can get wet." or "If a bug lands on you, it can sting you" or "If you're unconscious, you can't harm someone". They are shallow in that they fill gaps via straightforward inferences rather than more deliberative, imaginative processes. While they are simple, we can use them to build systems that understand and manipulate possibilities similar to the way we do—they can "just look" at what could have happened. For example, my system can spot an avoided knife attack, recognizing self-defense. More elaborative imaginative processes—such as those discussed by Pylyshyn (1973); Dehghani et al. (2008); Gerstenberg et al. (2017)—would be an interesting future extension of the ideas I present here.

Possibilities without probabilities Note also that, by design, these rules concern *possibilities* rather than *probabilities*. For the behavior I'm trying to model, for the kinds of hypotheticals I've described—"If only Romeo had learned that Juliet's death was a ruse!", "What if I lose my passport?"—our reasoning process seems to be based on evaluating *qualitative possibilities* more than degrees of certainty. While the odds can certainly *modulate* our reaction—a near-miss with danger has more impact than a distant worry—I find that for most of our regrets, anxieties, and hopes, it is the spectacle of the story, not its calculated likelihood, that moves us. We say "How tragic that Romeo died from mere miscommunication!", rather than "How tragic that Romeo died from mere miscommunication, when he had an 83% chance of surviving otherwise!"

Because the system knows what is possible and impossible, the modified story turns out differently. Presumption and censor rules enable the system to generate alternative scenarios and populate them with relevant details. In the next section, I will demonstrate how the system *analyzes* such hypothetical scenarios for their moral content.

2.2 We know what outcomes are harmful and why

> "What use are ideals if we cannot fit them to the universe as we find it?"
>
> — Claudia Gray

When the system generates a hypothetical scenario, the next challenge is to figure out what it's about. What are its relevant characteristics? Can we find harm in it, for example? And can we determine who—if anyone—is responsible? If we can appraise hypothetical scenarios, we can model behaviors and reactions—suspense, surprise, poignancy—that depend on knowing how the alternatives turn out.

For this moral reasoning domain, the relevant analysis consists of identifying and parsing harms. With this ability, we can identify moral reasons that depend on knowing what else could have happened. We can excuse a gunshot in a barroom fight as an act of self-defense. We can justify the pain of an injection by noting the noxious disease it prevents. We can condemn an idle bystander who doesn't do anything wrong, but who fails to help when helping is easy.

In this section, I introduce a computational model of our ability to identify and parse harms. The first challenge is that harms are so varied. We have an extensive vocabulary and conceptual framework for wrongdoing and its endless permutations: physical harm, social stigma, inconveniences, neglected duties, accidents, and so on. Another challenge is that moral knowledge is not a self-contained expertise, but a facet of the world: in the right context, *anything* can be infused with a moral dimension. For example, going to the beach on the weekend has no moral content—unless I had already promised to help a friend move, or the trip would disturb a delicate ecosystem, etc. And moral knowledge does not obviously depend on, say, knowledge of computers—yet in order to recognize what is harmful about *cyberbullying*[12], you need to understand a bit about what the internet is and how people use it. In this sense, there is no limit to

[12] See discussion on pg. 39.

what you might need to know in order to extract *all* the moral dimensions of an arbitrary situation. The knowledge required for a comprehensive moral reasoner is potentially as unbounded as commonsense knowledge itself.

Pattern elevation To make progress, I focus on two distinct kinds of competence. First is the ability to recognize and match particular moral events in the story—to notice particular specific harms such as 'death by poisoning' or 'theft of a vehicle'. Second is the ability to reason about the overlapping moral features that link different harms together—that arson and theft are both forms of property damage, for example, but that theft is sometimes reversible while arson is not. The principle is that while there's a multitude of *particular* harms, each potentially requiring specialized domain knowledge to understand, all harms are *explainable* as harmful by virtue of a comparatively small set of features[13]. Because when we can recognize the particular instances and decompose them into a shared vocabulary of features, we can compare them ("arson and theft are both forms of property damage") and contrast them ("theft is sometimes reversible while arson is usually not") and weigh them against each other.

The two competences are identification and parsing. The first scans for particular harms, flagging them; the second breaks down each harm into its constitutent parts so harms can be explained and compared. I call the combined process *pattern elevation*; it is the highly knowledge-intensive process of elaborating the implicit moral dimensions of a story.

Pattern nodes In my system, the basic pattern-matching units are called *pattern nodes*. These are patterns which represent narrowly-defined harms and which are matched against a story. For example, one pattern node might be "`xx stubs a toe`". The theory is that each pattern node matches a very specific sort of harm, capturing domain expertise. (The *domain-general* properties of a harm—features such as duration and reversibility, as well as taxonomical classification—are handled by a separate structure discussed in Section 2.3).

[13] Jackson (2016) discusses a similar approach in the field of philosophy.

For my purposes in this thesis, I implemented the matching mechanism using Genesis concept patterns (Winston and Holmes, 2018). These detect constellations of explicit and inferred events in a story, with the aid of commonsense inference rules—for more details, see Appendix B.

Pattern nodes, like the Genesis concept patterns they are built on, can involve multiple simultaneous events and causal links. For example, the particular harm of *supplanting a king* involves multiple events "`xx is yy's king, and yy kills xx leads to yy becoming king.`" Note also that the same event can match multiple pattern nodes, participating in many different harms. For example, a slap on the face could simultaneously be a social harm as well as a physical one.

Note that while I use Genesis concept patterns, in principle you could implement the matching component of this theory using a different mechanism. A specialized matcher might be required to identify certain harms—such as those that require extensive elaboration of the text, or those (like 'ostracizing') that take many concrete forms. Matchers might use sophisticated matching mechanisms, involving arbitrarily sophisticated data structures and relying on large knowledge bases and auxillary cognitive processing. Extending the system's capabilities with sophisticated matchers would be an interesting follow-up to this work.

Role-specific patterns identify *who* gets harmed Often, it is not enough just to know *whether* a harm has occurred: we need to capture additional structure such as the *participants* in the harm. We might like to know who specifically was harmed, and whether some specific person harmed them. For one thing, our value systems are characteristically *personal*. It matters, in other words, not just whether harm generically occurred, but whether it happened to the hero of the story or the villain, to a stranger or to a family member. More fundamentally, many thematic concepts like revenge and pyrrhic victory depend on keeping score between recurring characters. If A harms B and is harmed in return, that's revenge. If A harms B and then B harms C, that could be displaced aggression. If A harms B, and then X harms Y, with no further relation, there is no concept.

I address this need with *role-specific concept patterns*, which augment Genesis concept patterns with bindings for various roles like agent (person committing harm) or patient (person being harmed)[14]. Role-specific concept patterns provide a more sophisticated pattern matching apparatus, enabling the system to identify concepts that depend on who does what, such as revenge or self-defense. For a demonstration of how role-specific concept patterns are defined and used, see 4.5 Excusing self-defense.

Next: analyzing the quality of harms By identifying harms in alternative scenarios, the system models the kind of hypothetical reasoning we do when excusing preventive harms such as self-defense. Note, however, that this evaluation is rather coarse: as described so far, the system can detect *whether* particular harms occur and, using role-specific patterns, who the participants are. Yet the system lacks any further knowledge about these harms, including a sense of scale—the patterns for, say, a papercut, a murder, a stolen backpack, a broken treaty, and an outbreak of war, have no features to distinguish them. Without such features, the system has no way to distinguish wildly disproportionate acts of 'self defense', such as murdering someone to prevent them from tresspassing. The *quality* of a harm—its intensity and characteristics—matters. We compare how serious harms are; we see what features they have in common; we can explain what is harmful about them. The problem of modeling the *quality* of particular harms, so as to compare, contrast, and explain them, is the subject of the next section.

2.3 We compare outcomes qualitatively

To complete our ability to evaluate hypothetical scenarios, we must have ways to compare alternatives—to say that this outcome is worse than that one, or that this reward is commensurate with that favor. We must not only recognize harms but understand their characteristics. Comparison comes, in part, from knowing

[14]The odd terminology employing "agent" as the one who performs the action and "patient" as the one who receives the action is inherited from the theory of thematic roles in linguistics. See, for example, Saeed (2015).

what makes these harms harmful, what features they share with other harms, and what categories—such as property damage or bodily harm—they belong to.

The challenge is that our capacity for evaluation is subtle and highly complex: situations are so different from one another, it is hard to know what rubric to use. Real life is rich with particular details, and these details matter; they complicate our moral judgments, they provide conflicting cues, they introduce moral ambiguity, they make it difficult to identify which features matter most and almost impossible to decide that one situation is *strictly* better than another. What's more, our judgments often have a *qualitative* rather than *quantitative character*: we reason with loose orders of magnitude or by making analogies with earlier situations we've seen (Dehghani et al., 2008).

Only very rarely can we can compare situations along a single dimension, or using numerical measurements. In fact, I suspect that this only happens when humans specifically design artifical environments for doing so: In law, we have tort standards that codify in exacting detail which features matter (ignoring the rest!), and how much is owed in return (Graeber, 2012). In commerce, we buy products that have been designed to be uniform and comparable along a few easily-digested numerical measures such as size, weight, price, number of extra features, etc. In philosophy, we have thought experiments such as the (well-worn) trolley problems, which elicit our moral intuitions with impossibly sterile setups where only one choice is possible and the outcomes differ in exactly one way. Outside of such contrived situations, we must confront our moral dilemmas without simple straightforward numerical measures to guide our way.

How do we do it? To build the system described in this thesis, I developed a theory of how we humans make such evaluations. I focused specifically on our ability to make comparisons in complex situations[15] without falling back onto numerical measures. I implemented the theory as a program for making *qualitative moral comparisons*. The system can thus identify story concepts such as *escalating revenge* or *slap on the wrist*, adding nuance to mere retaliation by

[15]In this thesis, I use 'situation' as a general technical term, referring to the concept patterns that are detected when the system reads a story. These patterns can vary in scope from "theft", which comprises a single event, to "pyrrhic victory", which is a constellation of several events, to entire stories.

comparing who struck hardest. When added to a hypothetical reasoning capability, the system can identify hypothetical-based concepts such as *preventive harm* or *self-defense*, where you might excuse someone from committing a harm if it prevented a clearly greater harm from occurring. In this way, the system can supplement its understanding of the present situation by weighing it against hypothetical alternatives in much the same way humans do.

The qualities of human judgment The features of human moral reasoning we have just discussed form the basis for a theory of moral evaluation. Put precisely, in order to capture our human capacity to recognize concepts such as *escalating revenge* or *preventive harm*, we must have an evaluative mechanism which is:

1. Comparative—By design, its purpose is comparing one situation against another. This comparative context matters; for example, you might choose to consider different features of a situation depending on which situation you're comparing it against.

2. Multi-faceted—It must be sensitive to a rich, idiosyncratic, flexible universe of features. It's not enough for it to be able to compare some fixed family of features, as if comparison shopping.

3. Heuristic—Comparisons are qualitative, not just numerical. As a result, some comparisons are indeterminate; that is, sometimes we can't decide whether one situation is strictly better than another. And we don't demand absolute consistency, either: sometimes A is better than B, and B is better than C, yet C is better than A.

I present these criteria here in their final form. In Section 3.5, I survey alternatives and discuss why these features are especially compelling.

The moral lattice represents a qualitative value system In my system, I developed the *moral lattice* to encode knowledge of harms, their characteristics, and their relative magnitudes. A moral lattice is a kind of semantic network (Woods, 1975) for representing a particular person's value system. The moral

lattice[16] captures relative degrees of harm, such as "stubbing your toe is preferrable to breaking your leg". An example moral lattice is shown in Figure 2.1.

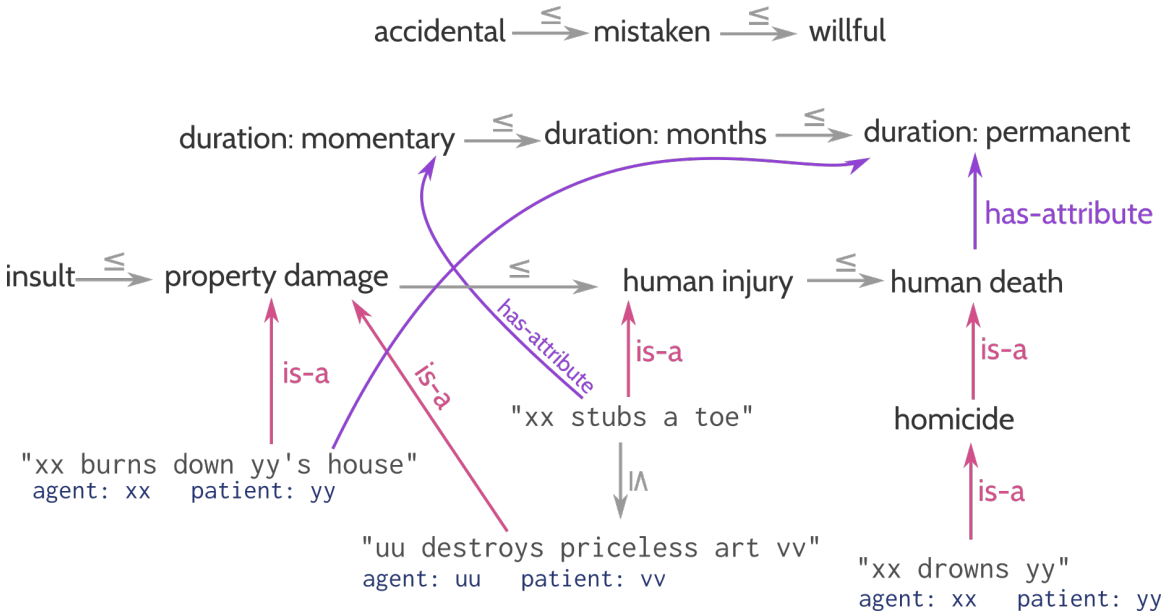


Figure 2.1: A *moral lattice* is a kind of semantic network for modeling a particular value system. It represents relative degree of harm using a system of nodes and directed, labeled edges. The endpoints of the network are patterns which can be matched against a story. Higher level nodes encode abstract properties such as *permanence*, *proximity*, etc.

The basic units of the moral lattice are *pattern nodes*, described previously in Section 2.2. These are patterns which can be matched against a story. For example, one pattern node might be "`xx stubs a toe`". These nodes are joined together with labeled directed edges. The simplest type of edge is an explicit declaration of relative harmfulness:

"`xx stubs a toe`" $\xrightarrow{\text{less-harmful-than}}$ "`xx breaks a leg`"

[16] I use the term 'lattice' informally, to evoke the picture of criss-crossing arrows between different strata representing different levels of abstraction. In mathematics, *lattice* is a technical term that carries additional assumptions, such as that arrows cannot form loops. My system is not a lattice in this formal sense, and indeed does contain loops, by design.

But more complex relationships are possible: besides pattern nodes, which are the basic units of the moral lattice, there are *feature nodes*. Feature nodes encode the abstract features of various harms, such as the fact that some harms are physical, some harms are more permanent than others, some harms befall close friends and others befall strangers, and so on. Feature nodes are linked to pattern nodes with edges such as `has-feature` or `is-a`. Feature nodes are linked to one another with similar links, such as `is-a` (creating a taxonomy, such as broken limbs and illnesses are types of physical harm), `has-feature` (which allows low-level concepts to inherit properties from their parent patterns), or `less-harmful-than` (which expresses general principles such as that property damage is, as a rule, less harmful than physical injury.)

In this way, the lattice links pattern nodes (which match harms in the story) to information about their general characteristics. By exposing the common abstract features of different particular harms, feature nodes allow harms to be compared. The web of links between feature nodes amount to a taxonomy or microtheory of harm. The possible library of node types and edge labels is deliberately open-ended, to allow for arbitrarily complex moral theories and extensions thereof. And the links are joined in a network rather than a tree, to escape the constraint of a single rigid hierarchy.

```
(def simple-concept-graph {:node-index {} :edges {}})

(let [e1 (katz-translate "xx breaks a treaty with yy")
      e2 (katz-translate "xx declares war on yy")]

   (def lattice
   (-> simple-concept-graph
        (add-node e1)
        (add-node e2)
        (add-edge e1 :less-harmful e2)))
   lattice)
```

Listing 1: A code fragment shows in detail how a one-off lattice is built. Note that events are declared and parsed from natural language sentences, and the lattice is built up by successively modifying the empty lattice, `simple-concept-graph`. The result is a simple moral lattice with two events—breaking a treaty and declaring war—with a single edge between them.

Given a story, the function of the moral lattice is first to identify the moral contents of the story—in our case, instances of harm such as, for instance, theft, kidnapping, property damage, insult, or injury. This is done by scanning the story for patterns matching the pattern nodes of the lattice. One moral patterns in the story are identified, they can be compared against each other by tracing appropriately-shaped paths through the lattice.

Path regular expressions define comparability So far, we have seen moral lattices, which supply a flexible language and syntax for describing value systems. To form a complete reasoning mechanism, we need *processes* for reading the lattice and extracting answers: given this web of features and their relationships, how do we decide if one situation is overall worse than another? Is one better along some dimensions and worse on others? Are they incomparable?

In my system, this function is fulfilled by *path regular expressions*, which are pattern-matching mechanisms that find specific kinds of path through the lattice. For example, you can use path regular expressions to define heuristics like these:

- If Situation A is connected to Situation B by a chain of explicit `less-harmful-than` links, then declare Situation A less harmful than Situation B.
- Otherwise, if Situation A has an attribute such as permanence or proximity which makes it worse, and Situation B has no similar property, then declare Situation A more harmful than Situation B.
- Otherwise, if Situations A and B are instances of more general types of harm, recursively check whether those supertypes are comparable.

Path regular expressions are rules for interpreting the moral lattice. They define specialized search routines that supply the moral lattice with semantic

meaning, explaining which constellations of arrows correspond to our conception of relative harm. In my view, it is really the path regular expressions that give the moral lattice its expressive power; with them, we can define matching rules for identifying harm, and precedence rules for resolving conflicting cues (e.g. if one situation involves serious harm to a person you've never met, and another involves temporary harm to a close family member, what specific details would make each situation the preferrable one?)

In implementation, path regular expressions are to semantic networks what regular expressions are to strings. They match directed paths made up of a specific sequence of labeled edges in the moral lattice. In the base case, a path regular expression consists of a single label to be matched, such as `less-harmful-than`. Expressions can be combined with various operators to form more complex expressions: *concatenation* requires that a collection of expressions occur in sequence within a contiguous directed path; *alternation* allows any one of a collection of expressions to be matched; *polymerization*[17] allows zero or more copies of an expression to be matched in sequence.

Each of these operators will be instantly recognizable to users of string regular expressions; however, there are several additional operators that are unique to the directed-graph data structure, developed to meet a particular moral-reasoning need. The *fork* operator is a kind of conjunction operator; it requires that all of the expressions in a collection have a match in the lattice starting from the same source node. The related *arc* operator requires that the matched paths share both the same source and the same terminal node. The operators *preclude* and *tail-check* act as filters on the matches accumulated so far: the unary operator *preclude* acts as negation, matching all nodes where the given expression does not occur. The operator *tail-check* takes a predicate as an additional argument; it filters for paths whose tail node satisfies the predicate. Finally, the unary *upstream* operator reverses the matching convention for directed arrows, following them "upstream" (in reverse). This kind of matching capability is useful for making two paths that "meet in the middle", and turns out to be necessary when matching any sort of `has-attribute`; see Figure 2.2.

[17] Note that this is the same idea as the Kleene star operator in string regular expressions

To summarize, the catalog has the following eight path regular expression operators:

- Concatenation, alternation (disjunction), and polymerization (Kleene star)
- Fork and arc (conjunction)
- Preclude (negation)
- Tail-check (node predicate test)
- Upstream (reverse-arrow mode)

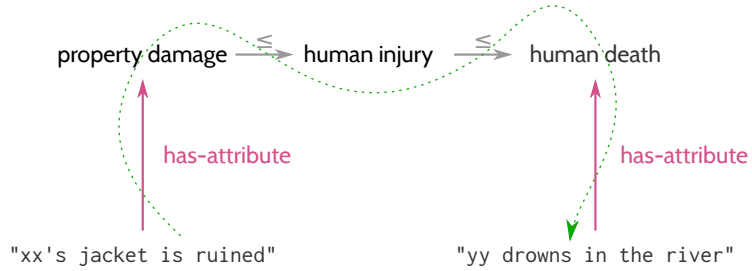


Figure 2.2: The *upstream* operator is required for finding paths between supertypes. In this simple excerpt of a lattice, the prospect of a ruined jacket—an instance of property damage—is compared against the prospect of a drowning person—an instance of loss of human life. To connect them appropriately, the path (dashed green arrow) must travel upstream along the second has-attribute edge.

For an example, see 4.1 Identifying harms that can be compared.

Escalating concept patterns require multi-factor search The moral lattice provides information about which harms are worse than which others. The Genesis story-understanding substrate[18] provides information about cause and effect. By integrating these sources of information, the system can identify high-level moral trajectories in a story, such as when a minor harm leads to a major retaliation. Trajectories can move in two possible directions: escalating action, in which minor harm causes major harm, and de-escalating action, in which major

[18] See Appendix B.

harm causes minor harm. Using these trajectories, the system can distinguish comparative concepts like *escalating revenge*, *slap on the wrist*, or *win the battle, but lose the war*.

The system finds moral trajectories using the following procedure[19]. It identifies the cause-and-effect relationships between events in the story ("perspective"[20]), as well as all the level-of-harm comparisons that can be made between story events with moral content—in other words, all effective paths in the lattice. The result is two lists—cause-and-effect pairs, and minor-major harm pairs. Using the two lists, it finds pairs of events that are *both* causally and morally linked. Depending on whether the cause or effect is a greater harm, it labels the pair as escalating or de-escalating action.

With role-specific patterns, we can construct even more sophisticated matches. Here is a definition of *escalating revenge*, in which three simultaneous conditions must be met:

1. Causal connection. Two events must be connected via a leads-to relationship in the story.
2. Escalating trajectory. The two events must constitute harms, where the final harm is greater than the initial harm.
3. Role reversal. The participants in the two harms must trade roles.

Our trajectory-finding algorithm `find-comparable-leads-to` does most of the work; to find escalating revenge, it is enough to check whether the trajectory is escalating and the harms have reversed roles. The fact that this check is straightforward suggests the effectiveness of our chosen representational scheme.

In this way, we can tersely define a search pattern for escalating revenge as a causal connection with increasing reciprocal harm. For an example, see 4.3 Finding disproportionate retaliation.

[19] Called `find-comparable-leads-to`.

[20] The Genesis system stores each story—along with reader context (such as commonsense background knowledge) and its own analysis (such as causal connections and themes)—in a `perspective`. See Appendix B for more details.

Extending the knowledge base To demonstrate the hypothetical reasoning capabilities in this thesis, I constructed a knowledge base consisting of several forms of specialized knowledge.

I built knowledge of possibilities and impossibilities (approximately a dozen presumption rules and censor rules, such as "`If xx falls in the water, xx can become wet`"), domain-specific harms (two dozen patterns, such as "`xx barks at yy; xx is the agent; yy is the patient.`"), and a particular qualitative value system (a persistent moral lattice). In that lattice, the feature nodes consist of a taxonomy of major harm types[21] (including property damage, social insult, mental anguish, bodily injury, death), five main attributes—permanence, reversibility, proximity, duration, intentionality[22]—as well as possible attribute values (e.g., the harm's *proximity* might be personal, or to a loved one, or to a stranger), the two dozen aforementioned pattern nodes, and the links between them. Links join feature nodes to feature nodes ("property damage is by default less harmful than personal injury..."), individual patterns to each other ("...but a paper cut is less harmful than the destruction of a famous painting"), and patterns to feature nodes ("xx stabs yy" is a form of murder).

The knowledge base is clearly of quite modest size; my main aim was to characterize the information we use, not to make a comprehensive encyclopedia. Nevertheless, it serves as a concrete instantiation of the hypothetical reasoning principles I discuss, and a demonstration of what kind of knowledge is required.

One important question about the feasibility of this work is how extensible it is. How much work does it take to add new information, and what are the limits of what it can represent?

As is typical with common sense knowledge bases, scaling is difficult in part because we don't know how to automate knowledge accumulation. The kinds of knowledge we want to represent—knowledge of possibilities and impossibilities, categories of harm and how people compare them—are frequently not

[21] Though the hierarchy isn't exclusive—recall that harms may belong to multiple or none of these categories.

[22] i.e., whether the harm was done on purpose, by accident, by mistake, or by omission. As in everyday language (Austin, 1956), accidents and mistakes are subtly different failures: 'By accident' applies to "I took your book with me by accident—I forgot it was in my bag", while 'By mistake' applies to "I took your book with me by mistake—I thought it was mine."

written down anywhere, and it can be laborious to elicit that information and convert it into a given representational framework.

Nevertheless, one of the contributions of my moral reasoning framework is a family of general, reusable characteristics that make it easier to introduce new harms. The principle is that while harms vary widely in their particulars and the amount of domain knowledge needed to understand their mechanisms, their *harmfulness* can be explained in terms of a small number of characteristics such as whether they cause physical injury, mental anguish, loss of social status, exhaustion, inconvenience, property damage, and so on. We can introduce new harms—even from a new domain—by relating them to these general shared characteristics.

Let me give an example. As societies change, new harms and new moral norms emerge. Before electronics were widespread, US laws about security were largely a matter of protecting property rights and punishing trespassing (Brandeis and Warren, 1890). It took the advent of tappable public phone lines, video cameras, and cloud storage to crystalize a new principle that *personal privacy* could be invaded even when no trespassing or property damage had taken place. For another example, before the internet existed, there was no such thing as *cyberbullying*.

Cyberbullying is a relatively recent form of harm. In order to include it in the moral lattice, it is first necessary to describe particular concrete forms of cyberbullying as they appear in stories. We add new pattern nodes: "`xx sends an anonymous rude text message to yy`", "`xx exposes yy's private information online`", "`xx harrasses yy in a video game`". These allow the pattern-matching apparatus to identify instances of cyberbullying. We can make the new category explicit by linking these patterns to a general category node `cyberbullying` via is-a relations.

Next, having included several concrete cases, we identify what characteristics make these cases harmful. We link them to their taxonomical types—many of the cyberbullying examples involve *mental anguish* and *loss of social status* rather than, say, *bodily injury*—and fill in the other characterstics such as *duration*, *permanence*, *reversibility*, and *proximity*.

In addition to listing characteristics explicitly, we can describe harms by analogy with others; description by analogy is often faster. Most harms, even in specialized domains, resemble others: Sending a rude text message is harmful for many of the reasons that sending a rude letter is harmful. Exposing private information online is similar to exposing private information in person or in a newspaper—although online publications have a uniquely wide audience and a long lifetime.

We make these connections explicit in the lattice. When we link the new harm "`xx exposes yy's private information online`" to a harm already in the knowledge base "`xx publishes yy's private information in a newspaper`", the new harm inherits all the characteristics of the old harm by default. Where the default is inappropriate, we repair it: we define the duration of "`xx exposes yy's private information online`" as permanent.

In this way, a small library of common features makes it easier to add new knowledge to the knowledge base, and moreover to expose what makes these harms harmful so that they can be compared, contrasted, and explained.

3 The structure of moral knowledge

3.1 Insights from cognitive science and philosophy

My system is a model of human moral reasoning. It embodies several ideas about how we humans think; some of these are grounded in cognitive science literature, others are grounded in particular philosophical traditions. In this section, I describe the components of the model and connect them to the concepts in the literature.

The first idea is that moral reasoning depends on what could have happened otherwise—what philosophers call the *counterfactual comparative account* (Hanna, 2016). In this view, our judgments of right and wrong depend on alternatives that never happened; if we only consider how things actually turn out, we miss key information. Accordingly, my system makes judgments by generating hypothetical scenarios, identifying their moral features, and comparing them against each other.

Generating scenarios To generate hypothetical scenarios, my system generates alternative *stories*[23]. In my framework, finding moral harms amounts to finding morally harmful *events* in stories. I based this characterization on Bradley (2012), who argued that while in everyday language, we talk about how activities, people, physical objects, and words can be harmful, in fact these are all implicit shorthand for the harmful *events* they take part in.

The generated stories provide key context for moral reasoning: the same event might be excusable in one scenario and inexcusable in another depending on what possibilities were available. By placing moral weight on possibilities, impossibilities, and constraints, I am employing the principle in deontic logic that our duties are based on what we're capable of (Von Wright, 1951): we might be excused if it was impossible to do better, and we might be condemned for doing nothing if we could have easily helped instead.

[23]That is, I represent hypothetical worlds using stories. See Appendix E for more details about the story representation, which is declarative rather than imperative.

Our moral judgments depend on our knowledge of what could have happened otherwise. In particular, via the deontic principle, they depend on knowledge of what choices we were capable of making and how those choices would have turned out. This knowledge can take a variety of forms depending on how it is used. For the behaviors I consider in this thesis, this knowledge is implemented as a collection of special rule types: *presumption rules* encode what might possibly happen, while *censor rules* encode what can't possibly happen. Presumption rules and censor rules offer one partial explanation of how we generate hypothetical scenarios and fill in their details. They complement the philosophers' careful analysis of which alternative scenarios matter morally (e.g., (Hanna, 2016)), providing a mechanistic explanation of how we construct these scenarios—what knowledge is required, how it is represented, and how it is manipulated.

Building these mechanisms helped expose just how effective we are at locating meaningful, enlightening context among innumerable alternatives. We humans have a keen sense of which possibilities are imminent and meaningful; the interplay between presumption and censor rules helps capture part of this phenomenon by controlling which scenarios are generated and which details are filled in.

Identifying morally-important features Once the scenarios have been generated, they must be analyzed for moral content. In my framework, the moral acceptability of a scenario is assessed by measuring how much 'harm' is in it, especially relative to nearby alternatives. This is, roughly, the *consequentialist approach* (Kamm, 2008) to determining moral acceptability: the moral choice is whichever one causes the least harm. (While the consequentialist approach does not account for every one of our moral intuitions (Kamm, 2008), it provides a good starting point for the harm-based judgments I model in this thesis.)

My system reasons about morality by identifying and comparing harms. I make a number of assumptions about how we human beings think about these harms, as follows:

1. Harms have internal structure. This makes them comparable (see below) and explanable.

2. Harms are unified (Hanna, 2016)— they are not just an arbitrary list of things that can happen to someone, but are based on a psychological theory of what constitutes harm (Ryff and Keyes, 1995).

3. Harms are qualitative, not quantitative. Degree of harm is not determined exclusively by numbers or numerical scores, but depends on other attributes.

4. Harms are one-sided, not two. That is, our moral calculus includes only *negative* concepts (harms), rather than both positive and negative concepts (harms and benefits).

   As a design decision, I have chosen to evaluate situations in terms of how much *harm* is in them, as opposed to measuring how much harm and benefit are in them, then weighing the harm and benefit against one another. I did this because I found that in practice, harms and benefits are rarely similar enough to be comparable. When a harm and benefit are comparable, they usually turn out to be the presence or absence of some specific feature—stubbing a toe versus not stubbing a toe, for example—in which case they are adequately captured by considering just the harm and its absence. Measuring only negative welfare leaves out some cases, but it provides a useful preliminary theory for the cases I consider in this thesis.

Harm is not the same as immorality (Lefkowitz, 2008). Consider surgery, which characteristically involves cutting or removing part of a person's body and therefore in this sense always involves *some* harm. For me, harm is a qualitative measure of badness, a scoring rubric. In some other amoral domain, we could assess and compare harmfulness like we would compare scores. For example, we could make decisions in a business domain by assessing the relative downsides of each choice.

3.2 What do harms all have in common?

> Hath not a Jew hands, organs, dimensions, senses, affections, passions? Fed with the same food, hurt with the same weapons, subject to the same diseases, healed by the same means, warmed and cooled by the same winter and summer as a Christian is? If you prick us, do we not bleed? If you tickle us, do we not laugh? If you poison us, do we not die?
>
> —*The Merchant of Venice*, III.i.49–61

As we have seen, the moral reasoning capabilities of my system depend on a knowledge base consisting of several forms of specialized knowledge.

The first form consists of knowledge of possibilities and impossibilities (approximately a dozen presumption rules and censor rules, such as `"If xx falls in the water, xx can become wet"`) (Section 2.1).

The second form consists of patterns for detecting domain-specific harms. These include two dozen patterns, such as "`xx stabs yy with a knife; xx is the agent; yy is the patient.`" or "`zz is a book. xx burns zz. zz belongs to yy. xx is not yy.`" (Section 2.2)

The third consists of a qualitative value system (embodied in a moral lattice) (Section 2.3). In particular, the value system includes:

1. A taxonomy of major harm types[24] (including property damage, social insult, mental anguish, bodily injury, death).

[24] Though the hierarchy isn't exclusive—recall that harms may belong to multiple or none of these categories.

2. Five main attributes of harm—permanence, reversibility, proximity, duration, intentionality[25]—as well as possible attribute values (e.g., the harm's *proximity* might be personal, or to a loved one, or to a stranger).

3. Knowledge of how to recognize particular harms in stories (the two dozen aforementioned patterns), and the characteristics of and relationships between those harms.

   These characteristics and relationships are encoded as links in a moral lattice, a type of semantic net. Some links encode relative harmfulness, either between general categories of harm ("property damage is by default less harmful than personal injury. . . ") or between particular patterns (". . . but a paper cut is less harmful than the destruction of a famous painting"). Other links associate harms with particular characteristics, such as their duration, or relate harms to one another ("xx stabs yy" is a form of murder).

The knowledge base is clearly of quite modest size; my main aim was to characterize the information we use, not to make a comprehensive encyclopedia. Nevertheless, it serves as a concrete instantiation of the hypothetical reasoning principles I discuss, and a demonstration of what kind of knowledge is required.

One important question about the feasibility of this work is how complete its representations are. That is, how complete is its ontology? Beyond the specific examples shown in this thesis, how much knowledge can it capture? What are its limitations, and what is out of scope?

How complete is the ontology? In my hypothetical reasoning framework, we make judgments by generating hypothetical scenarios, identifying their key features, and comparing those features qualitatively. For moral reasoning in particular, these features consist of various *harms*.

As I've discussed, one of my main assumptions is that *harm* involves a very small universe of concepts: there is a small number of ways in which something

[25] i.e., whether the harm was done on purpose, by accident, by mistake, or by omission. As in everyday language (Austin, 1956), accidents and mistakes are subtly different failures: 'By accident' applies to "I took your book with me by accident—I forgot it was in my bag", while 'By mistake' applies to "I took your book with me by mistake—I thought it was mine."

can be harmful—physically, mentally, socially, and so on. These categories arise from our human nature: what human flourishing consists of, and what detracts from it [26]. And so, despite the fact that the *particulars* of harm vary across cultures and time periods (see the next section for an example involving the relatively recent harm of *cyberbullying*), I assume that all human harms are united by a compact set of ideas of what harm *consists of*. I call this the *unity principle*:

> **The unity principle**: Harms are not arbitrary; what makes harms harmful is a small family of unchanging characteristics defined by human nature. These characteristics consist of, for example, physical harm, mental anguish, social abasement, property damage, and goal frustration.

This is a strong assumption, though one I believe is plausible. It implies that once we have defined our primitive set of harm types, we have exhausted all possibilities—the ontology is complete. We don't have to continue cataloguing new harms and retrofitting new features . We have a complete account of what makes harms harmful, and no additions are necessary for as long as human nature remains what it is.

This assumption can be grounded in some empirical results from the cognitive science literature. For example, Ryff and Keyes (1995) have found that human welfare (and its opposite, human harm) can be compactly described in terms of six factors, with similar results described by [Linton & Shaw, 2011]. And Young et al. (2007) have looked at specific features such as *deliberate intent*, probing how they affect our judgments of wrongdoing. In building my system, I have taken these empirical results as a template.

My ontology of harms accounts for the primitive categories of harm. It is *complete* in that—according to the unity principle, as supported by the literature—it arguably covers the small domain I set out to cover; any new harm will be explainable in terms of these basic categories, and will not require retrofitting. Unlike a general-purpose knowledge base like Cyc or OpenMind, my knowledge base has crisply defined scope, and within that scope its ontology appears

[26] I'm focusing here on *human* harms, as I do throughout this thesis, though of course the same idea applies to other species and their harms (Nussbaum, 2006).

to cover everything. General-purpose knowledge bases are intended to cover a wide range of human commonsense knowldge; in contrast, my system attempts only to capture the few features that make harms harmful.

3.3 How do we learn new moral knowledge?

A second important question about the feasibility of this work is how extensible its knowledge base. How much work does it take to add new information?

My system has both abstract and particular knowledge about harms. The abstract knowledge, as discussed above, consists of the characteristics that all harms have in common (general categories such as bodily injury). The particular knowledge consists of particular harms (such as stubbing a toe or getting stung by an insect), their features, and their relationships. Whereas the abstract knowledge consists of human universals, knowledge of particular harms requires domain-specific and culture-specific expertise; for each particular harm, the system needs a set of patterns for recognizing that harm in stories, and knowledge about what features it has.

As is typical with common sense knowledge bases, scaling is difficult in part because we don't have a way to automate knowledge accumulation. The kinds of knowledge we want to represent—knowledge of possibilities and impossibilities, categories of harm and how people compare them—are frequently not written down anywhere, and it can take hard work to elicit that information and convert it into a given representational framework.

Nevertheless, one of the contributions of my moral reasoning framework is a family of general, reusable characteristics that make it easier to introduce new harms. You don't have to define each one from scratch. The principle is that while harms vary widely in their particulars and the amount of domain knowledge needed to understand their mechanisms, their *harmfulness* can be explained in terms of a small number of characteristics such as whether they cause physical injury, mental anguish, loss of social status, exhaustion, inconvenience, property damage, and so on. We can introduce new harms—even from a new domain—by relating them to these general shared characteristics.

Let me give an example. As societies change, new harms and new moral norms emerge. Before electronics were widespread, US laws about security were largely a matter of protecting property rights and punishing trespassing (Brandeis and Warren, 1890). It took the advent of tappable public phone lines, video cameras, and cloud storage to crystalize a new principle that *personal privacy* could be invaded even when no trespassing or property damage had taken place. For another example, before the internet existed, there was no such thing as *cyberbullying*.

Cyberbullying is a relatively recent form of harm. In order to include it in the moral lattice, it is first necessary to describe particular concrete forms of cyberbullying as they appear in stories. We add new pattern nodes: "xx sends an anonymous rude text message to yy", "xx exposes yy's private information online", "xx harrasses yy in a video game". These allow the pattern-matching apparatus to identify instances of cyberbullying. We can make the new category explicit by linking these patterns to a general category node *cyberbullying* via is-a relations.

Next, having included several concrete cases, we identify what characteristics make these cases harmful. We link them to their taxonomical types—many of the cyberbullying examples involve *mental anguish* and *loss of social status* rather than, say, *bodily injury*—and fill in the other characterstics such as *duration*, *permanence*, *reversibility*, and *proximity*.

In addition to listing characteristics explicitly, we can describe harms by analogy with others; description by analogy is often faster. Most harms, even in specialized domains, resemble others: Sending a rude text message is harmful for many of the reasons that sending a rude letter is harmful. Exposing private information online is similar to exposing private information in person or in a newspaper—although online publications have a uniquely wide audience and a long lifetime.

We make these connections explicit in the lattice. When we link the new harm "xx exposes yy's private information online" to a harm already in the knowledge base "xx publishes yy's private information in a newspaper", the new harm inherits all the characteristics of the old harm by default. Where the default

is inappropriate, we repair it: we define the duration of "xx exposes yy's private information online" as permanent.

In this way, a small library of common features makes it easier to add new knowledge to the knowledge base, and moreover to expose what makes these harms harmful so that they can be compared, contrasted, and explained.

3.4 What about harms to society?

Most of the harms I've discussed so far have been harms between two specific people. But in many moral systems, it is possible to harm aggregations of people such as the poor or the taxpayers, or do injury to an abstraction such as a family, society, nature, one's honor, deities, or law and order. And some of these more abstract harms can involve people who are not direct participants in the actual event—such as when a child steals and is considered to have dishonored their parents.

My system can handle these abstract harms without any special modification. However, it is useful to delve into specifics to see how the system works in these cases.

Third-party harms The harms I've just described are what I call *third-party harms*, where the actions of individual people are considered to injure someone or some group not involved in those actions — a deity, for example, or a family, or a society. You can capture aspects of these with role-specific harms, such as `{:pattern "xx drinks alcohol", :agent "xx", :patient "deity"}`. These have the notable feature that the patient (victim of the harm) doesn't occur explicitly in the pattern — the victim must be inferred based on your societal knowledge.

For a person visiting a society for the first time, I'd expect these are among the hardest parts of the value system to *learn*, because the actual events can be harmless at face value until you've acquired a model in which the absent third party exists and can be harmed by these events.

On the other hand, the third-party nature of these harms does not make them difficult to *represent*. The type of harm is often a kind of social abasement (one of the dimensions of harm in my system, and the one involved specifically in honor, respect, rule-flouting, demotion, defacement, mockery, etc.)

By including these harms in the system, you can encode a value system in which "It is better that I murder this person to prevent a greater harm, namely an insult to the deity". My system's ability to build excuses out of hypotheticals works just as well for "I swatted their arm to prevent the bug from stinging them" (see 4.4 Excusing harms that prevent greater harms) as this kind of case.

Matching of aggregate groups and events One component of the system that I'd like to develop further is its ability to handle aggregate representations. An individual person withdrawing money from a bank is not harmful—but a run on the bank is. In some moral systems, there are large, often anonymous, aggregate participants—the poor, the taxpayers, the victims of automobile accidents, and so on. And there are aggregate events — repeated littering, habitual kindnesses, regular rituals. These have different character than their one-off counterparts, but I can represent them only thinly in the existing system, as atoms: "society", "taxpayers", "repeatedly withdrew money". The system doesn't have knowledge of parts and relations that would give it richer ability to explain how "taxpayers" relate to an individual taxpayer, or situate an individual withdrawing money within a run on a bank. As such, the pattern-matching part of my system depends, in these cases, a great deal on the level of abstraction at which events are represented.

In particular, in my framework, it is easy to represent and recognize how a single human sacrifice propitiates the gods, but representing the need for *repeated* human sacrifice is a bit fragile.

Honor as a fundamental type of harm Many types of harm are best described as harms to *honor*—one's public image, social status, or role in society. This is one of the primary types of harm in my system, alongside physical harms and emotional harms.

Harms to honor capture events such as an even accidental violation of a social taboo or of an oath. These harms may be powerful enough to cause violence or even self-inflicted injury or death.

For example, the notion of "Death before dishonor" is captured in my system by placing more value on certain forms of social abasement (losing face, looking bad, dishonor, shame) than on loss of human life. It can be your own life (as in dueling), or someone else's life (as in honor killings). The harmfulness can be modulated, or not, by whether the dishonorable act was intentional —- level of intent is one of the dimensions of harm in my system. As an example, if we place honor above human life, my system can model a person who chooses to face certain death in a duel over living in ignominity, in just the same way that it can model a person who chooses to ruin their jacket in order to save a drowning victim (4.6 Weighing outcomes when every choice is bad). The stakes are different, but the value judgement is analogous.

3.5 Why qualitative value systems?

> *"People say about every thing that it has a certain value. This is worth that. This coat, this sweater, this cup of coffee: each thing worth some quantity of money, or some number of other things—one coat, worth three sweaters, or so much money—as if that coat, suddenly appearing on the earth, contained somewhere inside itself an amount of value, like an inner soul [...] But what really determines the value of a coat? The coat's price comes from its history, the history of all the people who were involved in making it and selling it and all the particular relationships they had. And if we buy the coat, we, too, form relationships with all of those people, and yet we hide those relationships from our own awareness by pretending we live in a world where coats have no history but just fall down from heaven with prices marked inside. 'I like this coat,' we say, 'It's not expensive,' as if that were a fact about the* coat *and not the end of a story about all the people who made it and sold it."*

—Wallace Shawn, *The Fever*

My work in this thesis is inspired by a question: how is it that we readers find suspense, surprise, poignancy, danger, or luck in what we read? How do we recognize concepts like self-defense or preventive harm—concepts which fundamentally depend on harms that could have happened but didn't?

I have suggested that we can do these things because we are acutely sensitive to *hypothetical context*. We feel poignancy when we recognize a sad situation that could have easily turned out better. We recognize self-defense when a person commits one harm in order to prevent being harmed themselves.

One of my goals in this thesis is to explain how we represent and reason about hypothetical context. In order to do so, however, we must first explain what we *mean* when we say that a situation could have turned out "better", or turned out "worse". How do we capture our intuitive ideas of better and worse in a way that a computer can understand?

Qualitative comparisons rather than numerical Given that our goal is to model how humans assess relative harmfulness in ethical situations, an obvious approach is to use some kind of numerical scoring function to measure "how harmful" a scenario is. Indeed, this rough-and-ready approach has been taken by researchers such as the political scientist Goldstein (1992), whose *Goldstein index* defined a rudimentary quantitative scale for how combative or cooperative an international act is. On this scale, scores range from -10 (maximally combative) to +10 (maximally cooperative). For example, rejecting a proposal is mildly combative (-4.0), granting diplomatic recognition is rather cooperative (+5.4), and launching a military attack is maximally combative (-10.0). One of the chief advantages of the Goldstein index is that it provides a *uniform basis of comparison* for charting international relations in a quantitative way: a uniform numerical measure provides a good handhold on an otherwise heterogeneous saga of international incidents.

Numerical scores similarly form the basis for a prominent theory of ethics, *utilitarianism* (Bentham, 1948; Mill, 1859). Put very briefly, utilitarianism supposes that all forms of flourishing are uniformly measureable, that each person's

flourishing be treated as interchangeable, and that justice involves maximizing the quantity of flourishing. As a moral theory, utilitarianism makes equal treatment a primary value: no one person's flourishing is preferred over another's, and harms are all fungible, with utility as their common currency. Utilitarianism is appealing in that it is impersonally fair, and reduces moral problems to accounting ones. Like the Goldstein index, however, utilitarianism's numerical approach runs into some immediate objections[27]: *are* all forms of flourishing interchangeable? Does precise scorekeeping match our intuitions for how we make decisions in everyday life? Would it be morally acceptable—as fungibility suggests—to make a lot of people rather miserable so that one person will be transcendently happy? What about regularities and principles in our moral systems—ideals like prosociality, bodily autonomy, freedom of association? Don't we miss out on those higher-order structures when we flatten everything to a single number?

Numerical scores fail to capture certain aspects of our moral reasoning. Specifically, regarded as a representational framework, they lack internal structure, they carry a strong assumption that all harms are universally measurable and intercomparable, and they suggest a decimal-level exactness, linear rank, and systematicity that does not match our general intuitions about what we account for in everyday moral situations.

1. *Internal structure.* Numerical scores for international conflict do not expose what is similar about "expelling an organization (-4.9)" and "expelling a person (-5.0)", as social processes. They are similar because they both involve ousting an outsider group, and the resulting damage to social ties. Empowered by such explanations, we humans reason by analogies; we can anticipate generalizations and assign principled levels of harm to novel events such as "ousting a leader". We understand much more about human social dynamics, and therefore we understand much more about these harms than just a number representing how bad they are. For the

[27]Of course, utilitarians have answers to these objections; I cannot do justice to the full discussion here, and I am not presenting a knock-down argument. I am just reporting areas of the theory that are most strikingly unintuitive.

kinds of moral reasoning I study in this thesis, such internal structures are necessary.

2. *Intercomparability.* If numbers are different, one of them is always definitively greater, and by a precise amount. By assigning a single numerical score to each harm, we impose the assumption that all harms are intercomparable in this same way. But note that this is a terrifically strong assumption, and one that we do not necessarily need to make. Why assume that in our everyday life, we are always capable of deciding whether one harm is greater than another, and by exactly how much? Harms are heterogeneous and come in limitless varieties. To me, it seems extravagant to suppose we have a uniform scoring rubric that can meaningfully compare all possible harms along a single dimension of badness. And I doubt that such a score, even if you could compute it, would capture useful distinctions in everyday life. It seems to me much more plausible that sometimes our intuition fails to give us an exact answer as to which scenario is worse and by exactly how much. In this thesis, I use the *moral lattice* as a qualitative measure of relative harmfulness, providing one illustration of human-like reasoning which avoids a strong commitment to universal comparability.

3. *Exactness*. As with many rationalizing enterprises ranging from SAT scores to GDP scores, in moral reasoning the appeal of the numerical approach is that it discards complexity in favor of a quick summary. It produces a score against which whole populations can be unformly compared and linearly ranked. Indeed, one of the features of civil legal systems (e.g. in tort law) is that they do collapse the infinite variability of human experience to a simplified legal universe of discrete harms, features, and values. Such conventions make the law easier to administer uniformly, which is a kind of fairness (though see (Ensign et al., 2018)). On the other hand, even here, the numbers are a kind of summary—they are not the ingredients of moral reasoning, but their end products.

   There is a sense of arbitrariness to numerical representations: why a score of 1.0 instead of 1.5, or 2.0, or 100.0? Presumably in many applications, only relative scores matter, in which case the exact magnitudes are a su-

perfluous artifact of the representation. And numbers (at least real or rational numbers) are infinitely subdividable—do we mean to suppose that humans have or need the microscopically subtle distinctions that numerical systems supply? Finally, numbers are, by default, exactingly precise. Do we have a sense that our moral judgments inherit this kind of precise certainty?

There are quick-fix alternatives: Instead of flattening the moral universe into a single numerical factor, we might flatten it into a multidimensional array of *several* numerical factors. This evokes a principal components analysis approach, as with `word2vec`. Or instead of using single numbers, we could capture indeterminacy and coarse granularity with intervals or fuzziness (Zadeh et al., 1996).

Ultimately, however, I think each of these approaches mistakes the sort of processes I have aimed to model in this thesis. In my view, we use processes for generating narratives, identifying their features, re-prioritizing various features, forming analogies, proposing general principles, and considering hypotheticals. I believe that ultimately what happens in everyday moral reasoning involves narrative processing and debate that cannot be reduced to even coarse kinds of accounting.

Our current computing systems use numbers as a primitive type. As such, we naturally find it convenient map our representational schemes into numbers so that we can compute with them efficiently. But we should not be misled into thinking that these numbers preserve or reflect all of the structure of our representational schemes. Instead, we can use representations that abstract away from the underlying numbers. We can develop explicitly qualitative representations, such as the qualitative process theories of Forbus (1984a), or the lattice-based approach I have articulated in this chapter.

4 Moral reasoning demonstrations

I began by asking how we make judgments using hypothetical context. Using moral reasoning as a concrete case, I developed a set of hypothetical reasoning principles:

1. We think in terms of possibilities and impossibilities.
2. We connect details to general principles
3. We evaluate situations qualitatively

I built a computational model that replicates our ability to make hypothetical moral judgments and illustrates what kind of knowledge and processes such judgments might require. In the previous sections, I have described the particular tools I developed; these embody the high level hypothetical principles. *Presumption rules* fill in what-if scenarios, *pattern nodes* identify salient moral features, the *moral lattice* compares scenarios against one another.

In this section, I show how those tools work together. Putting my system through various demonstrations, I show how it can refer to hypothetical context when making moral judgments such as *excusing preventive harm*, *identifying self-defense*, and *evaluating counterfactual dilemmas*.

4.1 Identifying harms that can be compared

One of the basic functions of a moral lattice is to identify all harms in a story, via pattern matching, and then trace paths in the lattice to evalute how these harms compare to one another—more harmful, less harmful, equally harmful, or incomparable.

Note that with this subroutine, the system evaluates events *only* with respect to how much harm they contain, not when they happened or who they happened to. In the next examples, I'll show how we can augment information about the relative harmfulness of events with information about their *cause and effect* relationships and about who caused which harms.

Given a story to be read and a lattice encoding a particular value system, the system finds comparable harms according to the following procedure.

In the first stage, the system *finds which events are harmful and why*: For each pattern node in the lattice, scan the story for a match. As described in Chapter 2, patterns might include "`xx steals yy from zz`" or more complex structures. This process situates events in the story within their context in the lattice of harms.

In the second stage, the system *describes how harms relate to each other, where possible*: Using path regular expressions, search for directed paths between any pair of matched patterns. In practice, this connectivity search is made more efficient because the type restrictions on the path regular expression (such as "is-a" and "less-harmful-than") provide constraint.

In this way, the system produces pairs of events as output, each associated with a list of paths between them. These paths encode how the events are related to each other: less harmful, more harmful, equal, or comparable. As I will discuss in a later example, it is possible to define even finer distinctions, such as "*much* more harmful than".

Illustration: Comparing harms in *Macbeth* For the purpose of illustration, I use a handwritten English summary of Shakespeare's *Macbeth* to demonstrate how the lattice scans a story for moral contents and compares relative harmfulness.

> Shakespeare's *Macbeth* rendered in around eighty sentences.
>
> Scotland and England are countries. Dunsinane is a castle and Birnam Wood is a forest. Macbeth, Macduff, Malcolm, Donalbain, Lady Macbeth, Lady Macduff, Cawdor, and Duncan are persons. Lady Macbeth is Macbeth's wife. Lady Macduff is Macduff's wife. Lady Macbeth is evil and greedy. Duncan is the king, and Macbeth is Duncan's successor. Duncan is an enemy of Cawdor. Macbeth is brave. Macbeth defeats Cawdor. Duncan becomes happy because Macbeth defeats Cawdor. The witches are weird. The witches meet at night. The witches danced and chanted. Macbeth tells witches to speak. Macbeth talks with the witches. Witches predict that Birnam Wood will go to Dunsinane. The witches

predict that Macbeth will become Thane of Cawdor. The witches predict that Macbeth will become king. The witches astonish Macbeth. Duncan executes Cawdor because Cawdor is a traitor. Duncan rewarded Macbeth because Duncan became happy. Lady Macbeth wants Macbeth to become king. Macbeth is weak and vulnerable. Lady Macbeth persuades Macbeth to want to become the king because Lady Macbeth is greedy. Lady Macbeth wants to become queen. Macbeth loves Lady Macbeth. Macbeth wants to please lady Macbeth. Macbeth wants to become king because Lady Macbeth persuaded Macbeth to want to become the king. Lady Macbeth plots to murder the king with Macbeth. Macbeth invites Duncan to dinner. Duncan compliments Macbeth. Duncan goes to bed. Duncan's guards become drunk and sleep. In order to murder Duncan, Macbeth murders the guards, Macbeth enters the king's bedroom, and Macbeth stabs Duncan. Malcolm and Donalbain become afraid. Malcolm and Donalbain flee. Macbeth's murdering Duncan leads to Macduff's fleeing to England. In order to flee to England, Macduff rides to the coast and Macduff sails on a ship. Macduff's fleeing to England leads to Macbeth's murdering Lady Macduff. Macbeth hallucinates at a dinner. Lady Macbeth says he hallucinates often. Everyone leaves because Lady Macbeth tells everyone to leave. Macbeth's murdering Duncan leads to Lady Macbeth's becoming distraught. Lady Macbeth has bad dreams. Lady Macbeth thinks she has blood on her hands. Lady Macbeth tries to wash her hands. Lady Macbeth kills herself. Birnam Wood goes to Dunsinane. Macduff's army attacks Dunsinane. Macduff curses Macbeth. Macbeth refuses to surrender. Macduff kills Macbeth.

First, we need a moral lattice. For demonstration purposes, I'll define a minimal lattice here[28], including a few concepts such as murder and greed, as well as the relationships between them. We identify moral content by matching the pattern nodes of the lattice against the story. Note that matches may be complex, consisting of clusters of events, and events may match more than one pattern node, or match in more than one way. Here, the matcher successfully identifies Lady Macbeth's greed, as well as several murders, and compares their

[28]The system's knowledge base includes a larger lattice, but my aim is to show how lattices are built up in code and how the patterns are matched.

qualitative level of harmfulness. As an added detail, declarations such as `xx is a person` constrain `xx` to match only entities that are persons. Similar constraints restrict matches to animate beings, physical objects, etc.

Let's consider the simplest possible case. If, for example, we define harms of greed ("xx is greedy") and murder ("xx murders yy"), and link greed to murder with a single "less-harmful-than" link, the system proceeds as follows. First, it identifies four harmful events in this story:

- Lady Macbeth is greedy
- Macbeth murders Duncan's guards.
- Macbeth murders Duncan.
- Macbeth murders Lady Macduff.

Incidentally, note that each of these matches contains further information such as when it occurs in the story, what events it is causally related to, and who committed the harm—but for this procedure, that information is simply carried along, not used.

In the next stage, the system looks for paths between the matched events (as situated in the lattice). Because there is a directed "less-harmful-than" link between greed and murder, the system notes that "Lady Macbeth is greedy" is less harmful than each of the other events. Those events, because they are all harms of the exact same pattern, are considered equivalent by default.

The procedure returns these comparable events as pairs. For example:

```
(["Lady Macbeth is greedy." "Macbeth murders Duncan's guards."]
 ["Lady Macbeth is greedy." "Macbeth murders Duncan."]
 ["Lady Macbeth is greedy." "Macbeth murders Lady Macduff."])
```

By knowing which events in the story are harmful and how they qualitatively measure up against one another, the system has the raw material for further comparisons, analyses, and judgments. It can inspect how these harms are related causally (4.2 Identifying harms linked by cause and effect), account for who did what to whom (4.3 Finding disproportionate retaliation), and relate harms that happened to ones that didn't (4.4 Excusing harms that prevent greater harms).

4.2 Identifying harms linked by cause and effect

For many moral judgments, it is important to know not only which harms are worse than which others, but also which harm occurred first, or which one led to the other. My system understands these relationships, integrating causal knowledge and magnitude-of-harm knowlege. The moral lattice provides information about which harms are worse than which others. The Genesis story-understanding substrate[29] provides information about cause and effect. By integrating these sources of information, the system can identify high-level moral trajectories in a story, such as when a minor harm leads to a major retaliation.

Naturally, trajectories operate in one of two possible directions: escalating action, in which minor harm causes major harm, and de-escalating action, in which major harm causes minor harm. Using these trajectories, the system can distinguish comparative concepts like *escalating revenge*, *slap on the wrist*, or *win the battle, but lose the war*.

The system finds moral trajectories using the following procedure[30]. It identifies the cause-and-effect relationships between events in the story ("perspective"[31]), as well as all the level-of-harm comparisons that can be made between story events with moral content—in other words, all effective paths in the lattice. The result is two lists—cause-and-effect pairs, and minor-major harm pairs. Using the two lists, it finds pairs of events that are *both* causally and morally linked. Depending on whether the cause or effect is a greater harm, it labels the pair as escalating or de-escalating action.

With role-specific patterns, we can construct even more sophisticated matches. Here is a definition of *escalating revenge*, in which three simultaneous conditions must be met:

1. Causal connection. Two events must be connected via a leads-to relationship in the story.

[29] See Appendix B.

[30] Called `find-comparable-leads-to`.

[31] The Genesis system stores each story—along with reader context (such as commonsense background knowledge) and its own analysis (such as causal connections and themes)—in a `perspective`. See Appendix B for more details.

2. Escalating trajectory. The two events must constitute harms, where the final harm is greater than the initial harm.
3. Role reversal. The participants in the two harms must trade roles.

Our trajectory-finding algorithm `find-comparable-leads-to` does most of the work; to find escalating revenge, it is enough to check whether the trajectory is escalating and the harms have reversed roles. The fact that this check is straightforward suggests the effectiveness of our chosen representational scheme. In this new application, we can readily define the algorithm defined in terms of the primitives we have already assembled, as follows[32]:

```
(defn find-escalating-revenge

  (for [[[entity-1 entity-2] paths trajectory]
      (find-comparable-leads-to lattice perspective) ;; For every pair of
      ↪  causally+morally related events
      :when (= :escalating trajectory) ;; filtering for escalating
      ↪  trajectories only
    ]
    (let [harm-1 (:roles entity-1)  ;; pull out the role information about
    ↪  both harms
          harm-2 (:roles entity-2)]

    (when (and (same-entity? (:agent harm-1) (:patient harm-2)) ;; check for
    ↪  a role reversal
               (same-entity? (:agent harm-2) (:patient harm-1)))
    [(:agent harm-1) [entity-1 entity-2] paths trajectory [harm-1 harm-2]])
    ↪  ;; return the result
                                 )))
```

The `find-comparable-leads-to` subroutine performs the basic search; the main body of the algorithm augments the search loop by checking that the harms have well-defined roles and that the participants exchange roles. In this way, we can tersely define a search pattern for escalating revenge as a causal connection with increasing reciprocal harm.

[32] I have simplified the code presentation by eliminating some boilerplate list manipulations. The code is otherwise exactly as written.

For an example, see 4.3 Finding disproportionate retaliation, below.

4.3 Finding disproportionate retaliation

As a reprieve from the dark Shakespearan examples in earlier sections, I will illustrate the function of *escalating revenge* with the following story about dogs and cats:

> Start story titled "Retaliation". Nero, Oliver, and Felix are persons[33]. Nero barks at Oliver. Oliver claws Nero, presumably because Nero barked at Oliver. Because Nero barked at Oliver, Oliver clawed Felix.

Note that if you read over this minimal working example carefully, you can solve it by hand: intuitively, there is one instance of escalating revenge—when Nero barks at Oliver, and Oliver claws Nero in return. You can also notice a false lead, when Nero barks at Oliver, and Oliver lashes out at an innocent third party, Felix. Role-specific concept patterns help the system avoid mischaracterizing this case.

In reading the story, the system makes use of two patterns in the moral lattice: `xx barks at yy`, which is classified as a kind of insult, and `xx claws yy`, which is classified as a kind of bodily injury. Insults are, as a category, less harmful by default than injuries; hence barking is by default less harmful than clawing.

The `find-escalating-revenge` subroutine analyzes this story as follows: it identifies three separate harms—Nero barking at Oliver, Oliver clawing Nero, and Oliver clawing Felix. It identifies two causal connections—the barking causes both clawings—and two harm relations—barking is less harmful than clawing in both cases. Putting this information together, and checking for the appropriate role reversal, the system finds one instance of escalating revenge: the

[33] Side note: Actually, they are presumably *animals*, but for the Genesis knowledge base the *person* designator is a catch-all for all main characters or sentient beings, including talking animals that appear in fables, or countries that are anthropomorphized as having beliefs, desires, and actions.

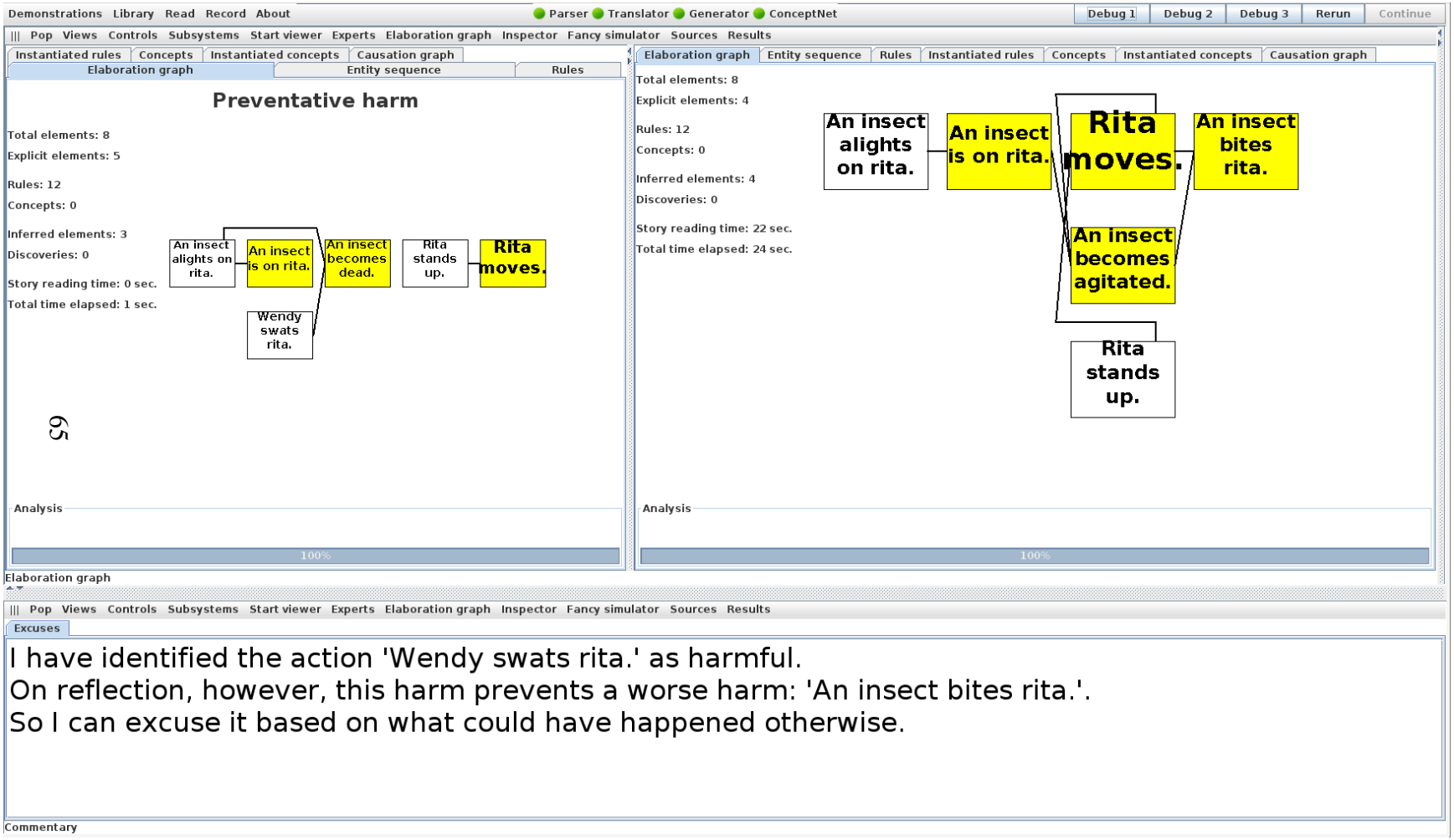
pair of linked events `["Nero barks at Oliver.", "Oliver claws Nero."]`. Note that the Nero-Oliver example was reported, while the Nero-Oliver-Felix near-miss was correctly excluded.

4.4 Excusing harms that prevent greater harms

There are many things that we do that seem morally objectionable on the face of it, while being morally excusable or even necessary in the appropriate context: a vaccine injection, while painful, serves an important preventive function; you might pull someone roughly in order to drag them out of danger; and you might attack someone in self-defense. Each of these explanations relies on hypothetical reasoning: reasoning about a harm that *could* have happened, but didn't. Sometimes the hypothetical context is so imminent and uncontroversial, we barely distinguish it as an act of imagination separate from filling in obvious commonsense details. In other cases, we take a greater imaginative leap from present context to imagined harm[34], and the excuses become flimsier[35]. In any case, we use our hypothetical reasoning apparatus as a fundamental part of judging right and wrong. Though we differ on the moral details and edge cases, the point is that this is a general competence that can be modeled.

In this section, I will show how my system can identify *preventive harms*—hypotheticals that excuse behavior—using the hypothetical and moral reasoning tools discussed so far. This serves as one piece of a demonstration of the breadth and flexibility of these tools for various forms of moral hypothetical question-answering. Figure 4.1 shows a worked example.

[34] See Schulman (2016) for a study of threat responses in conflict.
[35] See Austin (1971) for a survey of the conceptual landscape of excuses.



Rita and Wendy eat lunch. An insect alights on Rita. Wendy swats Rita. Rita stands up.

Figure 4.1: The program can use hypothetical reasoning to excuse harmful actions. In this particular scenario, a person swats at their friend to scare away a wasp. The system notes that swatting is harmful, identifies the dire possibility of a wasp sting, notes that the swat was preventive, and declares the harm excusable.

How can a computer program identify preventive harm? What we want is *cross-story comparison* of harms. In terms of the retelling paradigm (Appendix B), we can outline the desired procedure as follows:

1. Identify the central harm in the story ("original perspective").
2. Instantiate a new hypothetical perspective.
3. Copy the rule & concept knowledge into the new perspective.
4. Copy over the story, omitting the central harm.
5. Analyze the hypothetical story for its own emergent harm.
6. Attempt to compare the real and hypothetical harms against one another.
7. Report whether the avoided harm was worse.

Thus, the system begins by using its moral lattice to identify the harms of the story. If there is exactly one harm, the system labels it as central. If there are none, it gives up. If there are many, then by convention it picks a maximal harm[36]. In the case of our working example, the harm in the story is Wendy swatting Ria, an instance of temporary human injury.

Next, the system instantiates a new perspectiveand transfers over the common sense rules and concept patterns from the original reading. Third, it dismantles the original harm—omitting it from the story. In our example, it is sufficient to eliminate the swatting event in the second story. (In other cases, if the harm were a direct result or inference from other events, further dependency-directed backtracking would be necessary to completely remove the harm from the story.)

When the system analyzes the second, variant, story, a combination of knowledge in the form of censor and presumption rules lead it to conclude that the insect would have otherwise bitten Rita if it had not been swatted away. With two clear harms in view—the swat in the original story, and the insect bite in the hypothetical story—the system can attempt to compare the two. In some scenarios, the two harms will be incomparable, in which case the system will give a noncommital answer. However, in this case, according to the system's given moral lattice, an insect bite is more harmful (longer lasting, more distressing, potentially venomous) than a swat from a friend. Hence the system capably concludes

[36]That is, either the greatest harm in the story, or—because not all harms are comparable—a harm that is not outweighed by any other harm in the story.

that the swat was potentially excusable, given the greater harm that it prevented. It reports this result in the *Commentary* panel of the interface (Figure 4.1), using a combination of templated English text and language generated on-the-fly from its internal representations (quoted text).

In this way, the system is able to model our sensitivity to hypotheticals when reasoning about preventive harms. Though of course human judgments will vary as to whether an action is excusable or an imagined outcome is plausible, I argue that whatever judgments we make, we use the same machinery—a hypothetical apparatus like this one. What I suspect differ among people are our background knowledge and assumptions. I make these differences explicit in my system, as the value system is not hardcoded but rather plugged into the algorithm as a body of commonsense knowledge. To vary the moral framework of this system, it is sufficient to simply substitute a different set of commonsense knowledge, particularly an alternative moral lattice representing a different value system.

4.5 Excusing self-defense

Now that the system can identify specific harms in stories, it can scan hypothetical scenarios to see whether something harmful happens. For example, let's consider the bar-room fight from Chapter 2:

> George, Alex, and Martha are persons. Martha is George's spouse. Alex is George's lover. Martha and Alex despise each other. Martha encounters Alex and George at a bar. Martha yells at Alex. Alex brandishes a knife. Martha shoots Alex, then confronts George.

This is in part a story about harms that happen and that could have happened. We can define a set of patterns to capture our knowledge of what's harmful[37]:

```
{:pattern (katz-translate "xx shoots yy")
  :roles {:agent (variable "xx")
          :patient (variable "yy")}}
```

[37]Though my system has an accumulated non-volatile knowledge base of pattern nodes, as I discuss, a small one-off list will make a clearer example in this case.

```
{:pattern (katz-translate "xx stabs yy")
  :roles {:agent (variable "xx")
          :patient (variable "yy")}}
```

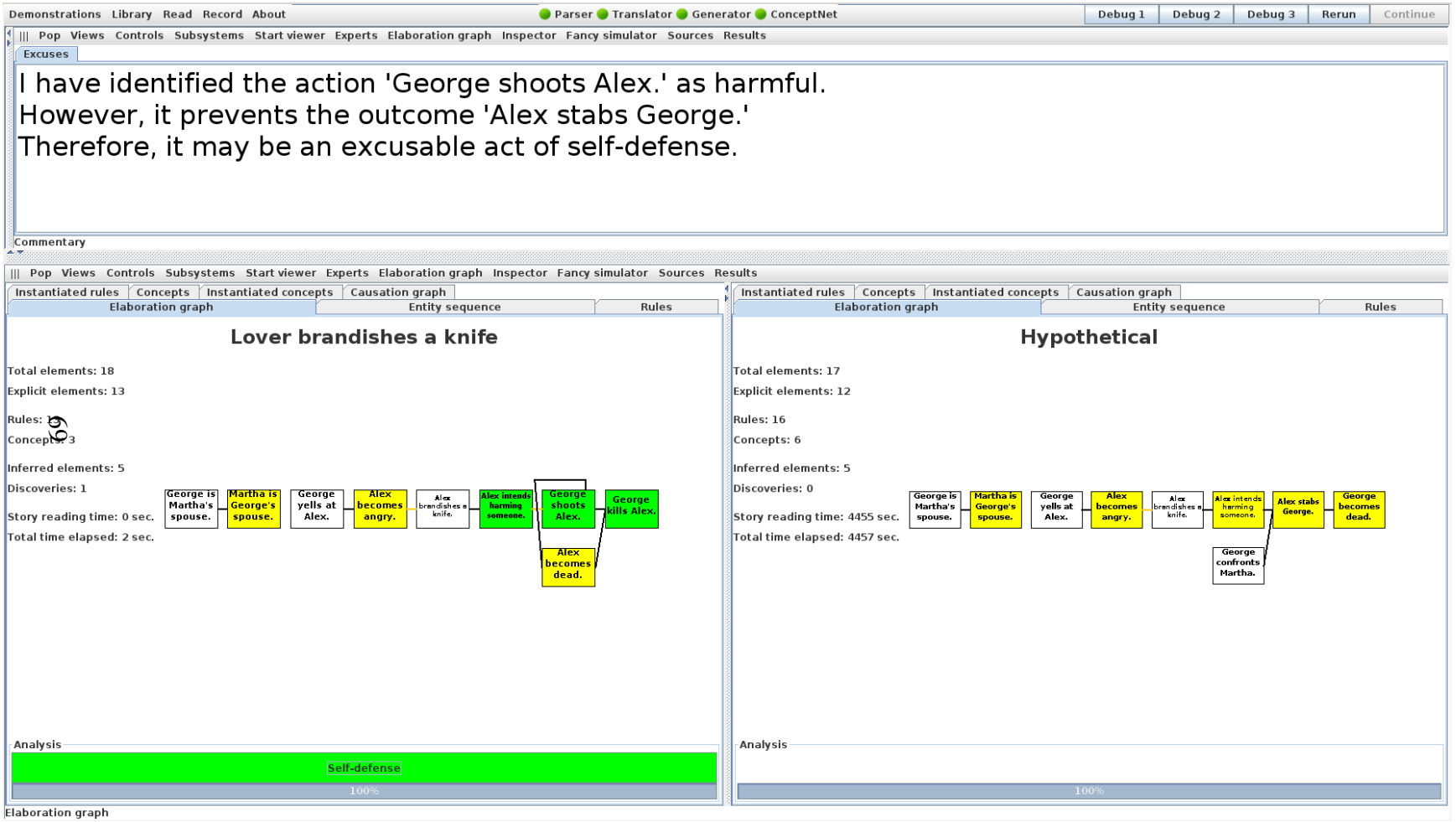
Listing 2: Role-specific concept patterns include bindings for specific roles such as agent and patient. The system uses these more sophisticated patterns to distinguish aggressors from victims, and to identify reciprocal harm in patterns such as escalating revenge.

When we ask the system "What would happen if Martha doesn't shoot Alex?", the system removes the event from the story and reads it anew. As a result of its presumption rules and censor rules, in the new scenario it finds that Martha becomes stabbed.

Using its pattern nodes, the system then identifies harm in the original scenario (Martha shoots Alex) and in the hypothetical scenario (Alex stabs Martha). The system notes the reciprocal roles: if Martha doesn't harm Alex in the original scenario, then Alex harms Martha instead. In this way, the system identifies three ingredients for a self-defense justification:

1. Initial wrongdoing: There is a harm in the original story
2. Revealed harm: When that harm is removed, the modified story has a new harm.
3. Reciprocal roles: The agent of the actual harm would have been the victim (patient) of the anticipated harm.

The general approach is encoded in a subroutine `find-self-defense!` which scans a story (or part of a story) for harm. It removes that harm, then reads the story anew. If it finds a harm in the new story, and the agent and patient have traded roles, it reports having found a potential self-defense excuse. A demonstration is shown in Figure 4.2.



George, Alex, and Martha are persons. George is Martha's spouse. Alex is Martha's lover. Alex and George despise each other. George encounters Alex and Martha at a bar. George yells at Alex. Alex brandishes a knife. George shoots Alex. George confronts Martha. The end.

Figure 4.2: Self-defense is a form of preventive harm in which the victim of the hypothetical harm is the perpetrator of the initial harm. In this tawdry story, George preempts a bar-room attack with an attack of his own.

The self-defense concept extends the preventive-harm concept as a special case. Going forward, you could easily create more nuanced refinements and computational theories of morality—to name just two, you might model degrees of *disproportionate self-defense* (when the preventive harm exceeds the harm it prevents) or form an ethical theory of *innocent victims* (innocent people who are harmed as part of an act of self-defense[38]).

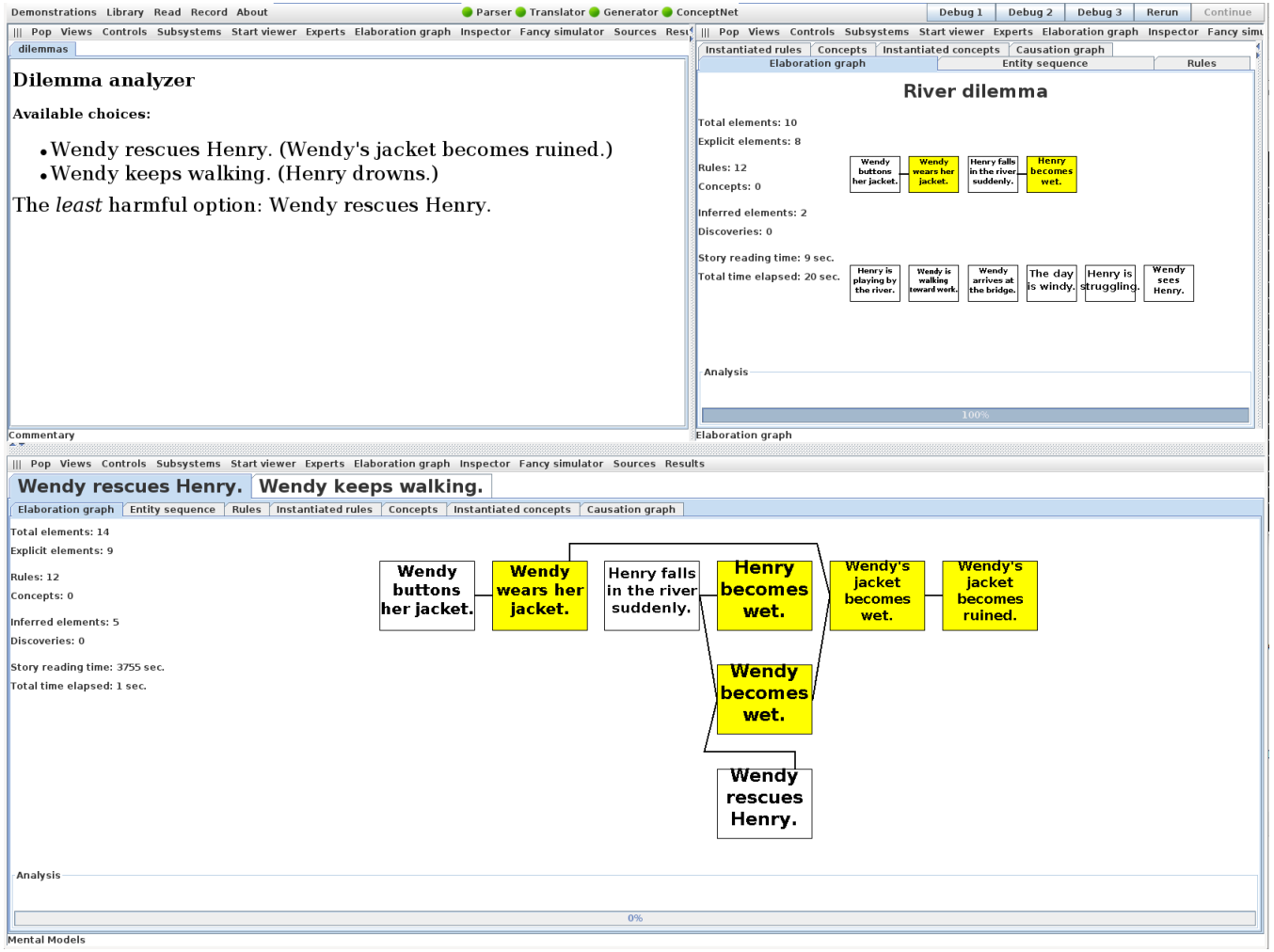
4.6 Weighing outcomes when every choice is bad

We have seen that the system can generate potential outcomes using presumptive knowledge, identify potential harms in those outcomes, and weigh those harms against one another. By integrating these capabilities, the system can perform *dilemma resolution*: when presented with a story and a list of possible choices, the system can select the best choice and justify that choice to a human user.

In the most straightforward case, for example, one of the scenarios is clearly less harmful than the rest, and the system reports it. Figure 4.3 provides an example of this clear-cut case with a story about sacrificing one's jacket to rescue a person from a river.

In most scenarios, however, there are conflicting cues: multiple harms, some greater and some lesser, as well as some harms that are not comparable at all. The result is a tangle of relationships, not all of which are interesting, actionable, or succinctly summarizable—such is the nature of moral dilemmas. However, using counterfactual questions (a form of hypothetical reasoning capability), it is possible to tease out insights and self-knowledge.

[38] See Kamm (2011).



Henry is playing by the river. Wendy is walking toward work. Wendy arrives at the bridge. The day is windy. Wendy buttons her jacket. Suddenly, Henry falls in the river. Henry is struggling. Wendy sees Henry.

Figure 4.3: In this scenario, the system evaluates whether Wendy ought to jump into the river to save Henry from drowning, potentially ruining her jacket. By envisioning the potential consequences of each choice, the system determines that rescuing Henry is least harmful.

4.7 Describing how changing factors change the verdict

In many moral situations, our judgments hinge on specific details. With counterfactual questions, we can prompt the system to analyze and explain which details make a difference to its own judgments. To answer a counterfactual question "How would the judgment be different if—", the system weighs the harmful consequences of each choice in the original and in the hypothetical scenarios, tabulating the results. The resulting table exposes the relative (comparative) harmfulness of the different outcomes, acting as a kind of shorthand signature. The signature indicates not only what choice is best in the original scenario and in the counterfactual scenario, but also the *rationale* for whether the counterfactual difference makes a difference. (Figure 4.4 shows an example.)

For example, depending on the circumstance, the system can express rationales such as:

- Although the counterfactual difference makes my original choice slightly more costly, it is still the best choice overall.
- Because the counterfactual difference makes the alternative even worse, the case for my original choice becomes even stronger.
- Because the counterfactual difference makes my original choice worse than the alternative, the counterfactual difference changes my decision.

In the rest of this section, I will demonstrate the range of answers my system is capable of constructing. Having decided that it is better for Wendy to rescue Henry from the river than not, the system can answer follow-up questions such as

- What if the jacket is expensive?
- What if Henry is a child?
- What if Wendy is a poor swimmer?

Although some questions result in the same course of action, the key idea is that the underlying rationale is different in each case.

With the first question ("What if the jacket is expensive?"), the system, appropriately enough, gives a scandalized response, which we might express as "How dare you! Although ruining an expensive jacket is worse than losing an ordinary jacket, neither is worse than losing a human life."[39] How does the system do this? Figure 4.7 shows the system's graphical output. The rubric is that if the original best choice (jumping in the river) remains the best choice in the counterfactual scenario *by a large magnitude*, despite having graver consequences, then the counterfactual question amounts to quibbling. The magnitude of the discrepancy is determined qualitatively in the network, via priority rules such as "loss of human life *greatly* outweighs property damage in general" (Figure 4.4).

	Original scenario	Hypothetical scenario	
Wendy rescues Henry	Jacket ruined	Expensive jacket ruined	$\nearrow$
Wendy keeps walking	Henry drowns	Henry drowns	$=$
	$\nearrow\nearrow$	$\nearrow\nearrow$	

Figure 4.4: "What if the jacket were expensive?" The system tabulates the harms that occur in each scenario and evaluates their relative badness where possible ($\nearrow$ indicates 'less harmful', $\nearrow\nearrow$ indicates 'much less harmful', $=$ indicates 'roughly comparable', etc.) The relative badness forms a signature that explains not only the preferred choice in the original and hypothetical scenarios, but the rationale for how the hypothetical affects this choice. Here, the system finds that it is still best to rescue Henry, even if slightly more costly–hence its 'How dare you!' response.

With the second question ("What if Henry is a child?"), the system gives an urgent response, which we might paraphrase as "All the more so—a child at risk is worse than an unspecified person at risk.". Accepting the system's child-friendly outlook for the sake of argument, the rubric is that if the original best choice remains the best choice in the counterfactual scenario while

[39]The *actual* output (Figure 4.7) is "How dare you! Granted, because 'The jacket is expensive', you pay a greater price when 'Wendy rescues Henry'. But that doesn't mean you'd be better off with 'Henry drowns'!". Note that although this response uses canned text, all of the single-quoted text is generated by the system and, more importantly, the system genuinely computes the rationale which is expressed here.

the *alternatives* become worse, then the original choice has heightened urgency. (Figure 4.5)

	Original scenario	Hypothetical scenario	
Wendy rescues Henry Wendy keeps walking	Jacket ruined Henry drowns	Jacket ruined & Child endangerment	= ↗
	↗ ↗	↗ ↗	

Figure 4.5: Tabulated results for the hypothetical question "What if Henry is a child?". The original choice to rescue Henry becomes relatively more urgent, as the alternative is even more harmful.

Note that while both this hypothetical ("What if Henry is a child?") and the previous hypothetical ("What if the jacket is expensive?") resulted in the same overall *decision* (to rescue Henry), the *explanation* for the choice was different in each case. For the expensive jacket, the explanation was that the decision to jump in the river became more costly but not by enough to matter. For the child swimmer, the explanation was that ignoring the swimmer became even more costly, strengthening the argument for jumping in the river. The same decision has different rationales; the system explains these rationales in terms of the table of relative harms it computes for each hypothetical question. The "shape" of the relative harms determines the structure of the rationale.

With the third question, ("What if Wendy is a poor swimmer?"), the system reverses its earlier judgment: "In that case, better keep walking. Losing Wendy and Henry is worse than losing Henry alone.". Taking the noninterventionist outlook for the same of argument, the rubric is that if the best choice in the counterfactual scenario is different than the best choice in the original scenario, we have discovered a difference that makes a difference (Figure 4.6).

Note that of course the choices have been artificially constrained; this scenario is unrealistic because the only two choices are risking one's life by jumping into a river, or blithely walking to work. Obviously, we do not live in a 'trolley problem' universe, and so there are often many more than two courses of action in any situation. The point is not that the system is reasoning about a realistic scenario, but that using its knowledge of the scenario, the system is capable of

constructing and articulating a human-like rationale for its decisions. It anticipates potential harms, weighs harms against each other, and tabulates the results in a compact form that captures its rationale. That is the key idea for this counterfactual reasoning capability.

	Original scenario	Hypothetical scenario	
Wendy rescues Henry	Jacket ruined	Both drown	↘ ↘
Wendy keeps walking	Henry drowns	Henry drowns	=
	↗ ↗	↘	

Figure 4.6: Tabulated results for the hypothetical question "What if Wendy is a poor swimmer?". Here is a difference that makes a difference, as it is no longer best to jump into the river.

Dilemma resolution is important because it showcases an important hypothetical reasoning competence: the ability to generate, compare, and contrast hypothetical outcomes. The system marshalls a significant volume of information comprising actual scenarios, imagined scenarios, and their counterfactual variants. Using the moral lattice framework, the system extracts and summarizes the key points relevant for decision-making. By tuning into the hypothetical context, the system is able to present lucid, human-readable moral evaluations like "Wendy should rescue Henry. Her jacket will get ruined, but that doesn't matter, even if the jacket is expensive!". And when—as in real life—the system encounters a scenario where there are many competing rationales and no clear winner, it can adroitly summarize these rationales for the human user.

I note that in future work, the system could use this kind of counterfactual reasoning to reflect on its own knowledge. By finding which differences make a difference, the system could discover and model how it makes its own inferences.

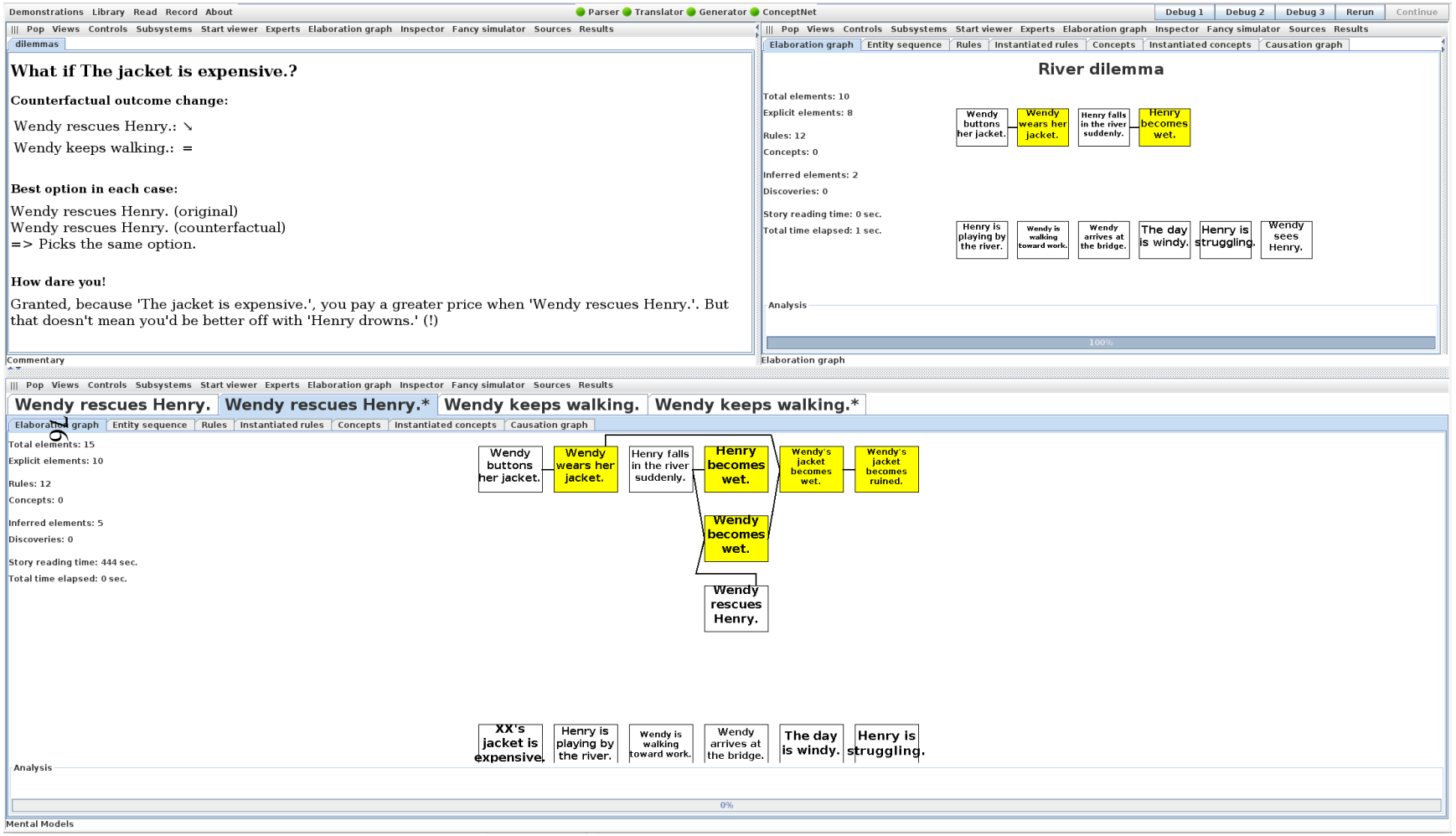


Figure 4.7: When reasoning counterfactually about a decision, the system weighs the harmful consequences of each choice in the original and hypothetical scenarios, tabulating the results. The resulting table acts as a kind of signature—an explainable rationale for keeping or changing the original choice.

5 Related work

In this thesis, I touch on many different aspects of AI. We've seen models of moral reasoning, control structures, cognitive architectures, search procedures, knowledge representations, constraint satisfaction, and expert systems—to name a few. We've also seen connections to other fields outside of AI, such as philosophy and cognitive science.

Here I acknowledge work in other fields that have provided empirical evidence, analytical theories, or computational models that relate to my work in this thesis.

Representations of internal processes I have argued for cognitive systems that represent not only the objective outer world, but also their own egocentric perceptions, internal states and processes. Gibson (2014) applied this principle to natural vision, developing the concept of affordances; Marr (1982) used it to explain perceptual artifacts such as optical illusions; and Sloman (2015b,a) proposed models of how we can imagine manipulating impossible figures. Where Gibson's affordances encoded visual opportunities for grasping, climbing, etc., I developed presumption rules to encode similar knowledge of opportunities for hypothetical reasoning in stories (such as where a harm might occur). Pylyshyn (1973) presented an insightful contrast between the properties of real images and mental images. Building on this contrast, I argued that stories are an effective representation for hypothetical scenarios: like mental images and unlike real images, stories enable you to include or omit details (such as shape, relative size, or color) as needed for a particular task. Collectively, although these lines of work differ from mine in that they deal with perception rather than narrative or reasoning, I followed their lead in explicitly representing egocentric concepts such as knowledge of possibilities ('presumption rules' and 'censor rules'), value judgments ('moral lattices'), and alternative scenarios (my 'perspectives'). More importantly, they convinced me that these internal representations need not be slavishly realistic. They may be abstract—only as detailed as the setting requires.

Representations for hypothetical reasoning A key subject of my thesis is how we represent the knowledge and processes of hypothetical reasoning. Rissland et al. (2005), whose work on case-based reasoning demonstrated how to apply knowledge engineering to the legal domain, inspired my own approach to moral reasoning in stories. My pattern nodes—which link particular events in stories to their supertypes in the moral lattice, thereby suggesting how to reason about them—can be seen as a similar case-based approach.

Winograd (1971) showed that language understanding is not just a matter of parsing; it is an integrated part of behavior and domain understanding. When Winograd's program carried on a conversation about blocks world, its vocabulary and conversational competence was a cornerstone of its blocks-world domain knowledge. Following Winograd's example, I developed my hypothetical reasoning program by focusing on enabling it to answer questions such as "Can we excuse this behavior on account of what could have happened?" or "What if the swimmer were a child instead?". That question-driven approach helped bring to light the concepts, representations, and processes necessary to reason hypothetically in general.

When representing knowledge of possibilities and impossibilities, I chose to build off of the Genesis rule-based framework. I developed pattern nodes to match themes in the story, and presumption and censor rules to add (or suppress) presumptive details to the story. This encoding of knowledge in terms of matched patterns and fired actions is similar to the logic-based work of researchers such as McCarthy (1980), Hayes (1979), and Singh (2005), although I do not use a formal deductive system. My focus on qualitative moral reasoning—in the form of the moral lattice and path regular expressions—builds on the qualitative reasoning work of Forbus (1984b), whom I discuss more below. Finally, the existence of knowledge bases such as Cyc and ConceptNet provided guidance and proof that building the moral knowlege base in this thesis was feasible and could be systemized. Gerstenberg et al. (2017); Gerstenberg and Stephan (2021) used detailed physical simulations to model our knowledge of possibility, impossibility, and causation-by-omission. In contrast, I focus on *narrative possibilities*, events that meaningfully could have occurred, regardless of the

probabilities. For example, we might feel the poignant idea that "Alas! Romeo died when he could have lived!", without computing a specific probability that, in some other story, Romeo might have lived.

I have argued that our interpretations are significantly modulated by context, specifically by the hypothetical alternatives we imagine. This has been approached in linguistics, with Gricean maxims (Grice, 1975) and so-called implicatures(Yang, 2016; Chomsky, 1965). Their central idea that we get much information from what people *don't* say directly inspired my approach to moral reasoning: when we make moral judgments, we account for the events that *don't* occur (for example, the harm that was avoided by an act of self-defense).

Building new representations In this thesis, I aimed for a system that could construct its own hypothetical scenarios. From the start, Karmiloff-Smith (1992) inspired me. This wonderful book describes representational redescription—how children change their mental representations as they grow up. With its catalog of diverse examples, it showed what we humans do. I thought then, and still do, that our ability to build new representations is a key part of our flexibility.

Magid et al. (2015) was a second touchstone. In their paper, they asked how we answer questions like "how do candy canes get their stripes?" There, the mystery is how we come up with answers that are *apt*, regardless of whether they're correct. How do we search so effectively? It was by trying to answer this question that I produced much of the machinery in 2.1. In particular, I developed the four meta-knowledge examples to explain how we avoid blind alleys and brute-force search.

More distant inspiration comes from Boden (1991)'s studies of computational creativity—computers producing works that are at least partly new. And of course, the idea of a machine that can improve itself goes back at least as far as Turing (1950).

Architecture Big flexible programs must be organized. They must use their knowledge effectively and add to it. They must choose goals and pursue them. Put simply, they need architectures.

Researchers such as Newell et al. (1959) developed some of the first cognitive architectures. They showed why architectures are necessary. Langley (2012) codified principles for building them. These principles—including a mandate to separate key theoretical tenets from implementation details—were a guiding star for my presentation in this thesis.

My program's architecture regulates its behavior using a simple model of emotion and deliberation. From Minsky (2006) and Singh (2005), I learned to see emotions as a regulatory mechanism—a fruitful idea, which I adopted to explain how we can rethink our initial moral judgments when in a more deliberative mode. Simon (1967) argued that such regulatory mechanisms help us decide what goals to prioritize. We use meta-knowledge —explicit knowledge of our own controls—to decide what to work on (Davis, 1980). Doyle (1980)'s dissertation in particular provided many striking insights about control structures in cognitive models. I use such control structures to explain how a system can modify its own search procedures so as to avoid wasted work.

Moral reasoning Moral problems are nothing new; I will point out some formative influences in my work. The philosopher Nagel (1979, Chapter 9) argued persuasively that not all values can be compared—a central tenet in my model. Von Wright (1951) introduced deontic logic—the moral calculus of permission and obligation—which resembles my moral evaluations in terms of possibility and impossibility. Though I deal with harms rather than promises, I share the principle that judgments are made by consulting alternative scenarios. Jackson (2016) articulated the importance of imaginary scenarios when making particular moral judgments. Austin (1956)'s charming article on excuses exposed some of the moral nuances that go unnoticed in everyday life. I often thought the article, with its careful catalogue of behavior, would've made a great start to an AI paper—I returned to it over and over while preparing this work. Saxe (2016) supplied corroborating evidence, highlighting neural mechanisms that modulate how we assign blame for accidents. Nagel et al. (1979) asked what role luck should play in our moral judgments. Magid and Schulz (2017) showed how we absorb secondhand values from others.

Studying morality requires a degree of self-awareness. The danger is chauvanism—claiming to have found the universal theory or the definitive ontology. Several writers kept me alert to this danger. D'Ignazio and Klein (2020) pointed out that our values always inflect what we study and how we study it. This is not a bug, but an invitation to self-reflect and to invite diverse perspectives. Broussard (2018) gave examples of narrow thinking in AI in particular. Their points were well-taken, and I strove to develop a moral architecture which enables alternative value systems instead of elevating one in particular. Finally, Wittgenstein (1979) surveyed the belief systems of many cultures, giving a rousing defense of moral pluralism against ethnocentrism.

Like Dehghani et al. (2008), I built a moral decision-making model which deploys an armamentarium of reasoning systems to make moral judgments. Their systems include order-of-magnitude estimates and reasoning by analogy as part of a reasoning-based framework. In contrast, mine include purely qualitative (magnitude-free) comparisons via the moral lattice framework. Analogical reasoning is, of course, a key element of flexible reasoning and would be an interesting future addition to the work I've done.

6 Toward the horizon

6.1 Learning large-scale moral knowledge

My system relies on knowledge of several forms—possibilities and impossibilities, story patterns, and how various harms relate to each other. Essentially all of that knowledge has been hand-coded. Such hand-coding was necessary at the start: I had a theory of what knowledge was required, and needed to supply my system with knowledge in that form. Unfortunately, it was of a rather specialized kind, namely things that meaningfully *could* happen in stories, but didn't; a characterization of why bumping your shin is bad; and a non-numerical explanation of why this is not quite as bad as breaking your leg. I could find no existing knowledge base that contained this knowledge, nor a data-driven approach that would allow me to mine the internet for it (the challenge: distill a corpus of events that meaningfully could have happened but didn't).

And so, I resorted to transcribing the knowledge myself, in the form the system needed. I wrote the presumption and censor rules the system used to generate alternative scenarios; I codified the attributes of harm into an ontology; and within that ontology, I captured and distinguished various everyday harms such as ruining a jacket, destroying a painting, or delivering an injection. With this knowledge, my system was able to model some human-like capabilities—it reasons about excuses, actions, and alternatives by referring to what could have otherwise happened.

That is, perhaps, enough to go beyond what was previously possible in some modest way. But where do we go from here? One of the main limitations of my system is the knowledge bottleneck: to be *broadly* useful, it must know a lot about harms and alternative scenarios. This knowledge is difficult to acquire automatically because it is the sort of information that every person knows, but which is not written down anywhere; moreover, it is highly-structured domain knowledge (including e.g. the taxonomy of harm and its various attributes) which makes it more difficult to glean from general-purpose knowledge bases.

I argued in 3.2 that some of this knowledge—specifically the ontology of harm itself—does not need much extension. I have argued, as expressed in the *unity principle*, that our psychological makeup dictates what is harmful to us, and that these types of harm (physical, emotional, social, etc.) suffice to classify harms across cultures, age groups, technological environments, and ethical systems. We differ in what causes us particular harm, but are united in what harm *consists of*.

While I've argued that the *ingredients* of harm are a relatively small set—physical and emotional and social harms, their duration, intent, reversibility, etc.—we can't reason about events in the world unless we know (or can infer) how much of each ingredient each event has. That set of harmful events is huge—potentially unbounded. There are, e.g., domain-specific harms and culture-specific harms which you'd miss unless you had the proper context. New technology and societal changes introduce new incarnations of harm (e.g. cyberbullying). But my system will be largely oblivious to these without a great deal of information about how to recognize them. The concern is that my system will only be a novelty, not a useful model of human moral reasoning, as long as it fails to explicitly account for the huge complexity, variety, and number of cases *genuine* human moral reasoning entails.

One response is to simply bite the bullet and admit that scaling up will be hard. Moral reasoning may indeed require a kind of general-purpose common sense, knowledge which might not be written down anywhere. And while the representations and processes I've built may pave the way for accumulating knowledge in the future, I didn't set out to build such a learning engine. Building that learning engine would likely entail dealing with a substantially separate set of questions and human competences than the ones I address in this thesis.

I think some of that response is reasonable: I chose this particular moral reasoning domain in part *because* I think that it is amenable to the kind of expert-system approach I took here, while data-driven methods might have a harder time. Moreover, I suspect children really do learn a lot by experiencing the rich physical world, by being teachable and being taught, and by having some useful initial frameworks and cognitive equipment to start with. For knowledge of that

specific kind, I'm skeptical that text is a good proxy for what we know. There may be some knowledge and some human competences that a machine (and maybe we humans) simply can't acquire that way, at least not without also having a great deal of mental architecture and/or rich physical and/or social learning environments already in place to support them. The actual required knowledge might be in our heads and in our experiences and nowhere else, in which case the only way to scale the system is to program that knowledge painstakingly. It is the standard problem with common sense knowledge bases in general; I would be surprised if all the structure and content of what we know can either be captured by reading or efficiently developed from a blank slate over sub-evolutionary timescales.

However, in this section, I would like to do more than throw up my hands and say that our ability to recognize specific harms depends on knowing a million facts (e.g. that a papercut is a type of injury, or that over-tight shoes can hurt) and is just one of those huge and unautomatable knowledge domains. Let me instead talk about what I think is possible.

1. Given an existing knowledge base of harms (such as the one I've built), you could design a system that reads stories and infers the relative moral weight of those harms from characters' actions. You would explicitly start with knowledge of harms, but not their comparative relationships, not the value system. This would be similar to the work in my Master's thesis (Holmes, 2017): The purpose of that program was to learn, based on a series of stories about a character, what their value system was. In detail, the system knew about what means characters could use to achieve various goals, and what moral constraints—such as avoiding harm, obeying laws, avoiding stealing—they might be following, so by examining their actions (and, of course, the actions they could have taken but didn't), the system inferred what constraints (what moral values) the characters might be following. It was interesting because a lot of the comparisons and reasoning were about situations that didn't happen (e.g. a person who doesn't steal when given the opportunity).

2. Although, as I argued in 3.3, instances of harm are infinitely varied in their particulars and precise in their internal structure, I do think you could plausibly guess categories of harm by aligning stories with an existing general-purpose common sense knowledge base. Physical harms are especially concrete: if you have knowledge of body parts, ailments, and attributes like sharp/heavy/hot/itchy, then you can presumably capture a large variety of physical harms that occur. That is, you could capture the more common and straightforward forms of physical harm, with coverage probably following a kind of Zipf's law. I'd expect there'd still be gaps in stories where a lot of context/knowledge is implied (e.g. plumbing: "Why did the person in the shower shout when someone else flushed the toilet?"), and that you would not necessarily get precise structure-matching such as who did what to whom, but you'd at least get a general sense that the story contains some type of physical harm.

3. As for emotional harms and social harms, I imagine something like sentiment analysis could give you a handhold: if you read stories where characters emotional reactions (and sources of those reactions) are stated explicitly or easily inferred, you can then deduce emotional harms and social harms by proxy: If a character feels shame or left out, for example, you can infer that something has socially harmed them. And, for that matter, you'd have another route to inferring physical harms whenever someone reacts with "Ouch" or a grimace etc. You wouldn't have to code every last one of these cues by hand, either; if off-the-shelf sentiment analysis doesn't quite fit, you could compare words by similarity to a few anchor words. This kind of emotion-tracking might be an interesting approach toward inferring type of harm.

4. The Genesis system already uses some amount of domain knowledge and word similarity when doing its matches—it's more than mere string matching. For example, Genesis uses WordNet threads when identifying concept patterns. When you're checking whether `"xx kills yy"`, for example, the pattern matcher also accepts hyponyms (such as "murders" or

"stabs"). The pattern matchers I've defined use WordNet categories (such as verb.communication[40] and verb.contact[41]) to define a broad scope for what they will match.

There are a few other tricks as well, including outsourcing commonsense knowledge from ConceptNet so that not all inferences or permutations of ways of saying things have to be coded by hand (Williams et al., 2017; Williams, 2017; Winston and Holmes, 2018).

5. Beyond type of harm, there's the question of how, in a particular value system, harms compare against one another. I think there, too, you might be able to leverage some existing knowledge base. A few knowledge bases track "relative badness", loosely construed. For example, ConceptNet (Speer et al., 2008) has a notion of "badness" that was one of the primary dimensions of variability discovered by principal components analysis. While you potentially lose the explanatory qualitative dimension ("this is worse than that because of these features...") and the cultural specificity of the value system (many crowdsourced knowledge bases are a tacit average or amalgam of everyone's knowledge system), you would at least be able to deploy some of the moral-trajectory mechanisms I developed for my thesis to a greater range of story outcomes.

Let me conclude this way. On reflection, I believe that the learning problem is not about automating knowledge acquisition—it's about coverage. It's about making sure that a system that purports to model moral reasoning *actually grapples with* the large variety of moral complexity that exists, rather than just planting a flag there.

My aim with the discussion in this section is to take a first pass at grappling with that complexity and with what it would mean for my system to be capable of handling examples at scale. In particular, I make the case that this thesis work is far more than a custom built system that does only what it was explicitly programmed to do. I've suggested a few extension projects—gist-based harm

[40] Suitable for many social harms.
[41] Suitable many physical harms.

detection, for example—to show that my work continues to be useful even when I've stopped hand-coding knowledge. I think one strength of these projects is that they rely on existing text-processing techniques and knowledge bases; they don't require any speculative innovations. So while I don't think these projects would teach us more about *how moral reasoning works*, I think they're a key part of the discussion: I developed my system's representations and processes by working through a small number of carefully-chosen examples in detail. Extension projects like these provide an roadmap for how this story-based moral reasoning system could begin to grapple with all the stories of harm that are out there in the world.

7 Contributions

I've argued that we understand the world in part because we are attuned to its hypothetical context. By working through particular examples in moral reasoning and human problem solving, I have aimed to explain in general how our hypothetical faculties work: what kinds of knowledge we need, how that knowledge is represented, and what processes manipulate it. This work has led to several key insights, which I distill here.

Hypothetical reasoning is fundamental to human intelligence The central idea of this thesis is that hypothetical context supplies an important part of our understanding. Examples drawn from everyday life show how pervasive it is: we able to fill in commonsense gaps, excuse preventive harms, flag unsolvable problems, imagine the future, and find poignance in what we read, only because we understand what could and can't possibly happen.

I have argued that, in particular, we rely on hypothetical context when making moral judgments: we excuse harms that are committed in self-defense, for example, and we condemn bystanders who do nothing when helping out is easy. I built a computational model that replicates this behavior and illustrates what knowledge and processes such judgments might require. The particular tools—*presumption rules* to fill in what-if scenarios, *pattern nodes* to identify salient moral features, the *moral lattice* to compare scenarios against one another—helped suggest what knowledge and processes are required for hypothetical reasoning in general. In particular, I identified a specialized knowledge requirement—knowledge of possibilities and impossibilities—which encodes what we understand about what could have happened otherwise. Then, because there are countless scenarios that *could* have happened otherwise, we require procedures to regulate *when* we consider hypotheticals, to *fill in* hypothetical scenarios with the appropriate level of detail, and to *evaluate* them using adjustable criteria.

We know about possibilities, impossibilities, and constraints We understand much more than what's explicitly in front of us. We understand what could have been. We understand the available alternatives so distinctly that they ground our basic moral judgments and elicit powerful emotional responses—suspense, surprise, poignancy, and the rest. We don't say: 'What I've just read *might be* poignant', or 'The incoming knife blow *might be* hazardous'. We say: just *look* at what could have happened—as if it's right there for everyone to behold.

This understanding requires *incisive* knowledge of alternative possibilities: not an undifferentiated haze of possibilities, but cogent outcomes that can ground our judgments. In my moral reasoning system, this knowledge takes the form of commonsense rules. *Presumption rules* encode what might happen (such as, if you fall into the water, you may drown), while *censor rules* encode what can't happen (such as, if a person is unconscious, they cannot harm you). I demonstrated that by using such rules, we can build systems that understand and manipulate possibilities similar to the way we do—they can "just look" at what could have happened. For example, my system can recognize an act of self-defense by spotting the harm that it precluded and excuse a painful slap that averts an even more painful insect bite.

Naturally, this approach has its limitations: first, not everything we know about possibilities and impossibilities fits neatly into a rule-based framework. Second, my rules contain no information about likelihood: in my system, I don't attach probabilities to outcomes, or even rank them in order of likelihood.[42]For some applications, such features are indispensible, and so the system's knowledge of possibilities and impossibilities will eventually need to be expanded to include them. But this approach, as it stands, provides useful insight: It shows concretely what you can do when you represent not just the explicit story as-is, but also the what-ifs that surround it. It demonstrates why we need specialized

[42]This was partly due to behavior I was modeling. For the kinds of hypotheticals I've described—"If only Romeo had learned that Juliet's death was a ruse!", "What if I lose my passport?"—our reasoning process seems to be based on evaluating *qualitative possibilities* more than degrees of certainty. While the odds can certainly *modulate* our reaction—a near-miss with danger has more impact than a distant worry—the narrative features of our regrets, anxieties, and hopes seem to play a more fundamental role than the precise probabilities.

commonsense knowledge about possiblities and impossibilities. And it suggests one possible instantiation.

The surprise is that, although alternative possibilities are always provisional, we nonetheless find them imminent and tangible enough to support our basic understanding of the world.

We connect detailed harms to general principles How do we identify harms in stories? In this work, I divide the process into two stages: recognition and analysis. In the recognition stage, we use pattern-matching to identify precise, domain-specific kinds of harm such as breaking a leg. In the analysis stage, we situate that pattern in context, identifying its particular type (as physical, emotional, social harm, or a combination), attributes (such as duration, proximity, and reversibility), and relationships to other harms (such as inheritance or analogical relationships).

I developed the *moral lattice* representation to capture both aspects of these processes. A moral lattice is a semantic net consisting of pattern nodes (which can be used to scan the story for particular harms) linked to their characteristic features and to other harms. In this way, the precise, domain-specific patterns are embedded in a network that describes their abstract moral features and relationships to other patterns.

The first key idea about the moral lattice representation is that it uses a *compact moral vocabulary*. I argue that although there are innumerable *particular* instances of harm—and these will vary among individuals and cultures and over time—there are only a small number of explanations for *how* something is harmful—as a physical harm, an emotional harm, a social harm, and so on. I argue that these general explanations of harm are human universal. They comprise the aspects of harm that we all understand, even if we do not relate to their particular application, and they comprise a relatively fixed set.

The second key idea is that as a result of this separation between knowledge-intensive, culturally dependent pattern knowledge and a compact vocabulary of common moral features, this system's moral knowledge base scales more easily: to add new types of harm, you add new particular patterns into the existing moral

framework.

We compare harms qualitatively I have described how humans reason about real-world harms, grappling with a great deal of complexity, indefiniteness, and nuance.

1. First, our judgments depend on context and particular features. Harms are a constellation of events in a story, and so we must pay attention to small details, which can often modulate our judgments. Moral judgments can hinge on the specific context and the specific details in the story.
2. Second, our judgments can be indeterminate. Harms often have so few features in common that we have little basis for comparing them outright—is it a worse harm to break your arm or total your car? In some contexts, the choice may become clear. In most, the choice is underdetermined.
3. Third, our judgments are rarely unilateral. Even when harms share enough features to be comparable, we must often deal with conflicting cues. For example, is it a worse harm to destroy a priceless painting or to stub your toe? According to one line of reasoning, human injury is generally worse than property damage. According to another, a stubbed toe is easily healed, while the destruction of a painting is irreversible. Cues like these pull us in many directions; sometimes we can decide on a winning argument, often we can't. In that case, we simply describe the considerations.
4. Fourth, our judgments are qualitative and interpretable. When we reason about harms in real life, we are not reducing everything to numerical scores, where harms are comparable precisely and univerally. We are not even comparing vectors of numbers or numbers with uncertainty attached. Instead, we extract key features, reason about them, and weigh them against each other. Our judgments are qualitative and interpretable. While numerical factors—orders of magnitude, probabilities, etc.—can certainly *modulate* these judgments, it is the spectacle of the story that determines them.

By defining comparability in terms of *paths through the moral lattice*, I was able to capture these qualitative attributes:

1. Judgements depend on context and particular features—indeed, paths are not required to give consistent comparisons. Frequently, there are loops $A \leq B \leq C \leq A$, which can sometimes be resolved one way or another depending on context. In short, the moral lattice is *globally inconsistent but locally useful*.

2. Judgements can be indeterminate—indeed, if one harm has too few features in common with another, there will be no paths between them. Instead, harms with features in common cluster into islands of comparability.

3. Judgements are rarely unilateral—between two nodes, there may be multiple paths, and paths in opposite directions. Each path corresponds to a particular argument for why one harm is greater or less than another. Sometimes precedence rules can determine a clear winner; the rest of the time, the system can describe the lines of argument by summarizing the paths in each direction.

4. Judgements are qualitative—because the elements of the path are edges in a semantic net, they are interpretable. Paths can be understood as *articulated arguments* for why one harm is less than another.

This system computes moral hypotheticals

I began this work with many questions about moral reasoning: How can hypotheticals—those subjective, provisional imaginings—ground our moral judgments? When we read a story, how do we identify what the moral harms are? And when we think about concepts like revenge and reparation, how do we ever weigh one harm against another?

These are the interesting questions that underlie our everyday ability to grasp moral problems. Hypotheticals are interesting because they reveal just how much of our understanding is built out of the knowledge we carry with us: our understanding does not end at what explicitly happens either in real life, on the page, or in a dataset; instead, a great deal comes from what we *know* could or could not have happened otherwise. And moral reasoning is interesting because it is ubiquitous, complex, and urgent—I suspect we all have common mental equipment for moral reasoning, the way we all have lungs or kidneys or language areas in the brain. I suspect we are all—regardless of age or culture—able to reason with a vocabulary of help and harm, dominance and fairness, precepts and magnitudes. In this view, our considerable differences in moral judgment therefore amount to different frameworks defining *which* things are harmful, how harmful they are, and in what ways. From a scientific perspective, it is useful to understand both this general system and our individual frameworks. From an engineering perspective, it is useful for the systems we build to reason articulately and transparently about the choices they make. Hence why these questions about moral hypotheticals are so crucial.

My approach to answering these questions was to build a computational model of our moral-hypothetical behavior. This computational approach helped sharpen the questions to be answered (What knowledge do we need? How should it be represented? What processes will manipulate it?) and helped expose surprising difficulties (We humans spot would-be harms in stories without apparent effort—how can a computer spot such possibilities among innumerable others?)

Accordingly, I gave an account of what our hypothetical moral reasoning

behavior consists of, complete with new computational concepts (presumptive knowledge, pattern elevation, qualitative harm comparison, etc.) I also built a running system to demonstrate how it might work; this system included particular concrete instantiations of these high-level concepts (in the form of presumption and censor rules, the moral lattice, path regular expressions, etc.) In this way, I provided some preliminary answers to the questions that inspired this work[43].

The system I built exhibits a brand-new competence—the ability to perform certain moral hypothetical computations in much the same way we do. The system is valuable in part because it provides a collection of computational proposals of how we work and a working demonstration of their effectiveness.

Now, although I grapple with questions about moral reasoning in this work, it is fundamentally a thesis in computer science and artificial intelligence. Of course, a computer science thesis about moral reasoning risks being reductive—flattening human complexity into an algorithm, munging human variation into a one-size-fits-all model, canonizing cultural bias with a veneer of objectivity. If we want to take a computational approach, we must guard against these risks. On the other hand, if we want to grapple with systems of great complexity, nuance, and richness, a computer is an indispensible tool. And if we want expressive power to describe complex ideas precisely, nothing beats a computer language. Used properly, the computer can *expand* the level of complexity we can handle, can *sharpen* the ideas we can express, and can *challenge* our assumptions about what is easy or universal or what goes without saying. It is in this spirit that I've undertaken this work.

In my view, this work complements the extensive existing literature on moral reasoning in philosophy and cognitive science. With my computational approach, I have brought certain practical questions into focus—how do we think of *meaningful* hypotheticals? How do we, mechanistically, compare one harm against another? What do we *know* about what could happen and how do we represent and manipulate that knowledge?

[43] Of course, in the process, I accumulated many new questions that await future investigation.

I have also introduced several new ideas about possible mechanisms: where utilitiarianism reduces all harms to universally intercomparable numbers, for example, I show one way to reason *purely qualitatively* about *features* of harms and show why some harms might reasonably be incomparable. I show how you can combine moral knowledge and cause-effect knowledge to reason about harm trajectories. And I show what is difficult about picking out the harmful particulars in a story and how pattern elevation provides one solution.

In this work, computational mechanisms like these suggest a few new directions for investigation. My hope is that the computational concepts in this work will provide some fresh ideas and a perspective that will be useful and stimulating for cognitive science and moral philosophy.

I have built a system that computes moral hypotheticals in much the same way that we do. In this way, I have helped to shed light on an important piece of how we humans operate: how we reason about better and worse outcomes, how we dream up possibilities and shut down impossibilities, how we are moved by reactions like suspense, surprise, and poignancy.

The key insight is that we are able to understand in large part because we are able to imagine. In particular, we are able to understand *moral reasons* in large part because we can imagine *moral outcomes*. We see, we anticipate, we judge, we compare, we consider. Imagination is at the very heart.

Appendix

A Why stories?

In this work, I have chosen to represent hypothetical scenarios using *stories*. Stories are an effective representation because a story can be amended to suppress or expand detail as needed: If certain details are irrelevant—such as the shape of a block, the specific material a knife is made of, or the exact dimensions of a room—you can omit them from the story. Conversely, if certain details become important, you can expand the description to include them. For example, to explain why someone has a bump on the head after looking for a lost earring, you might expand the description of the search to include particulars such as searching under the bed.

Why is the level of detail essential? When answering a question about a hypothetical scenario, you must imagine its particulars. For any given scenario, you know much more than you need. Only some of the details will be useful, and then only if they are represented with an appropriate level of granularity. Fill in too much, and you'll waste time; fill in too little, and you'll be unable to answer. Stories, then, enable you to calibrate the level of detail because a story is a kind of *declarative description*. A list of sentences is not a model of the world, but a blueprint for building a model out of explicit text and the reader's commonsense background. The story's description can be amended to suppress or expand detail as needed. This declarative approach is in contrast to a more imperative approach, such as a typical physics engine, in which details such as weight, shape, size, and material must be specified in full even when irrelevant to the task at hand.

B The retelling paradigm

As described in Appendix B, the Genesis Story-Understanding System is a general-purpose model of human intelligence built on a foundation of telling, under-

standing, aligning, and recombining stories, broadly construed. In order to direct its faculties toward imagining hypothetical alternatives, I developed an approach I call the *retelling paradigm*. The idea is to repurpose Genesis's story contexts, normally used to capture different reader perspectives of a story—such as different political views of the same story—or to let the system analyze its own problem-solving apparatus (Winston, 2018). I use them to store stories representing hypothetical alternatives.

In effect, I treat a hypothetical scenario as a person's subjective view of the present situation: all of the knowledge used to construct the imaginary details are made explicit as commonsense knowledge, the same way that subjective biases are made explicit in different perspectives. Knowledge of how to fill in imaginary details, and which ones to fill in, is a type of knowledge.

To achieve this, I implement a workflow in which the system first reads some initial story into a default perspective("perspectiveone"). The story is usually accompanied by an explicit declaration of commonsense knowledge (in the form of Genesis rules) and thematic concept patterns, which elaborate the story and drive thematic analysis. After the story is read—or perhaps while it is in progress—we can introduce a hypothetical question, something like "How might this character solve this problem?" or "Can you [the system] excuse the wrongdoing in the story, based on some greater harm that it helps to prevent?" To answer the question, the system populates a new perspective("perspectivetwo") with a fresh copy of the original story. Based on the specific question being asked, the system removes elements from the copy story and fills in new imagined details. Then the system can deploy the full Genesis analytical apparatus for identifying themes, making comparisons and analogies, introducing new commonsense information, etc., in order to analyze the modified story. A hierarchy of detail allows explicit information and deductive information to fill in gaps before imagined hypothetical details are filled in. In particular, the system can deploy the mechanisms developed in this thesis for identifying moral content, elaborating imagined details, spinning out additional hypothetical alternatives in their own perspectives, etc.

When implementing the retelling paradigm, one obstacle I encountered is

that the Genesis's boxes-and-wires architecture was originally hardwired and static; it was not initially developed to spin out arbitrary new perspectivesor attach new cognitive monitors during runtime. Solving this problem involved some creative work, the end product of which was a `sleeper-box`. A `sleeper-box` is a self-contained box whose instances can be dynamically created and attached to the Genesis system during runtime. Each one is equipped with a specific event trigger (such as "finished analyzing the story") and subroutine to run when triggered; hence they endow Genesis with dynamically modifiable event-listening.

I report the crystalized subroutine here for clarity and future reference for others who are interested in building similar design patterns:

```
;; CLOJURE
(defn box-sleeper
      "Create a wiredbox that waits until the model (perspective) has
      ↪  finished reading the
       story, then executes the function (fn-after this signal)."
      ([fn-after]
       (box-sleeper (genesis.GenesisGetters/getMentalModel1) fn-after))
      ([model fn-after]
       (let [FROM_MENTAL_MODEL "tmp-input"
             box
             (proxy [specialBoxes.MultiFunctionBox] []
                    (getName [] "sleeper-box")
                    (getPortName [] FROM_MENTAL_MODEL)
                    (process [#^Signals.BetterSignal signal]
                                 (fn-after this signal)
                                 ))]
            ;; STORY PROCESSOR -> BOX
            (wire-link! model box
                (storyProcessor.StoryProcessor/COMPLETE_STORY_ANALYSIS_PORT)
                       ↪  FROM_MENTAL_MODEL)
            (.addSignalProcessor (connections.Connections/getPorts box)
            ↪  FROM_MENTAL_MODEL "process")
            box
            )))
```

```java
// JAVA INTEROP
package specialBoxes;
import connections.WiredBox;
/**
 * This class serves as a base for creating WiredBoxes in clojure
 * dynamically.
 * @author rlm
 */
public interface MultiFunctionBox extends WiredBox {
        public String getName();

        public Object process(Object ignore);
        public Object process0(Object ignore);
        public Object process1(Object ignore);
        public Object process2(Object ignore);
        public Object process3(Object ignore);
        public Object process4(Object ignore);
        // --- snip. etc.
        public Object process29(Object ignore);
}
```

Listing 3: A `sleeper-box` enables dynamically-created event listeners in the Genesis system, a key feature for hypothetical reasoning. The Java portion defines a `MultiFunctionBox`, a bare-bones box with many wiredbox function stubs but no implementation details. In Clojure, the powerful `proxy` method dynamically constructs `MultiFunctionBoxes` with specific implementation details, wires them into the Genesis boxes-and-wires network, and destructs them once they've fired.

In summary, the retelling paradigm is the idea that in order to analyze hypothetical variants of a story, you can simply treat the hypothetical variants as stories of their own which are produced by modifying the original. Within this paradigm, you treat knowledge of which imaginary details to fill in as its own kind of specialized expertise, encoded as commonsense knowledge. As part of my system's implementation, I developed `sleeper-boxes` to dynamically create new story contexts and populate them with hypothetical scenarios.

The Genesis Story-Understanding System

Here, I briefly review the components of the Genesis story-understanding system upon which this thesis is built. For a more detailed background of the motivations for and capabilities of the Genesis system, I refer the reader to our foundational documents (Winston, 2011, 2012a,b; Winston and Holmes, 2018).

The *Genesis story-understanding system*, developed by Patrick Winston's research group at MIT, is a computational architecture which models how humans understand and tell stories. The overall vision of the group is that human intelligence is uniquely distinguished from the intelligence of other species by our ability to use and manipulate deeply nested symbolic descriptions—stories, broadly construed—and that if we are to understand and model human intelligence, we must understand and model the mechanisms that enable these story understanding capabilities.

In a typical use case, Genesis reads a text file of about twenty to thirty lines containing a story in simple English. After the story is processed, it is shunted to a variety of different agents—an arrangement inspired by propagation networks (Radul, 2009). These agents are specialized for a variety of intelligent tasks such as identifying questions, representing movement through space, accumulating knowledge, forming models of what characters know, judging tone, and forming a self model, among many others. The agents dispatch on the incoming sentences, construct their own internal representations, and pass messages to one another throughout this cognitive system so as to assemble a detailed comprehension of the story along many different dimensions.

One of the more fundamental representations used by Genesis is the *elaboration graph* (Figure 1). The elaboration graph is a structured representation of the elements of the story as a directed graph, with arrows indicating causal connections or inferential connections. (Causal connections include, for example, a problem precipitating a character's response; inferential connections include, for example, the conclusion that if a character is in the kitchen of a house, the character is consequently also in the house.)

Such commonsense connections are essential to understanding the story in

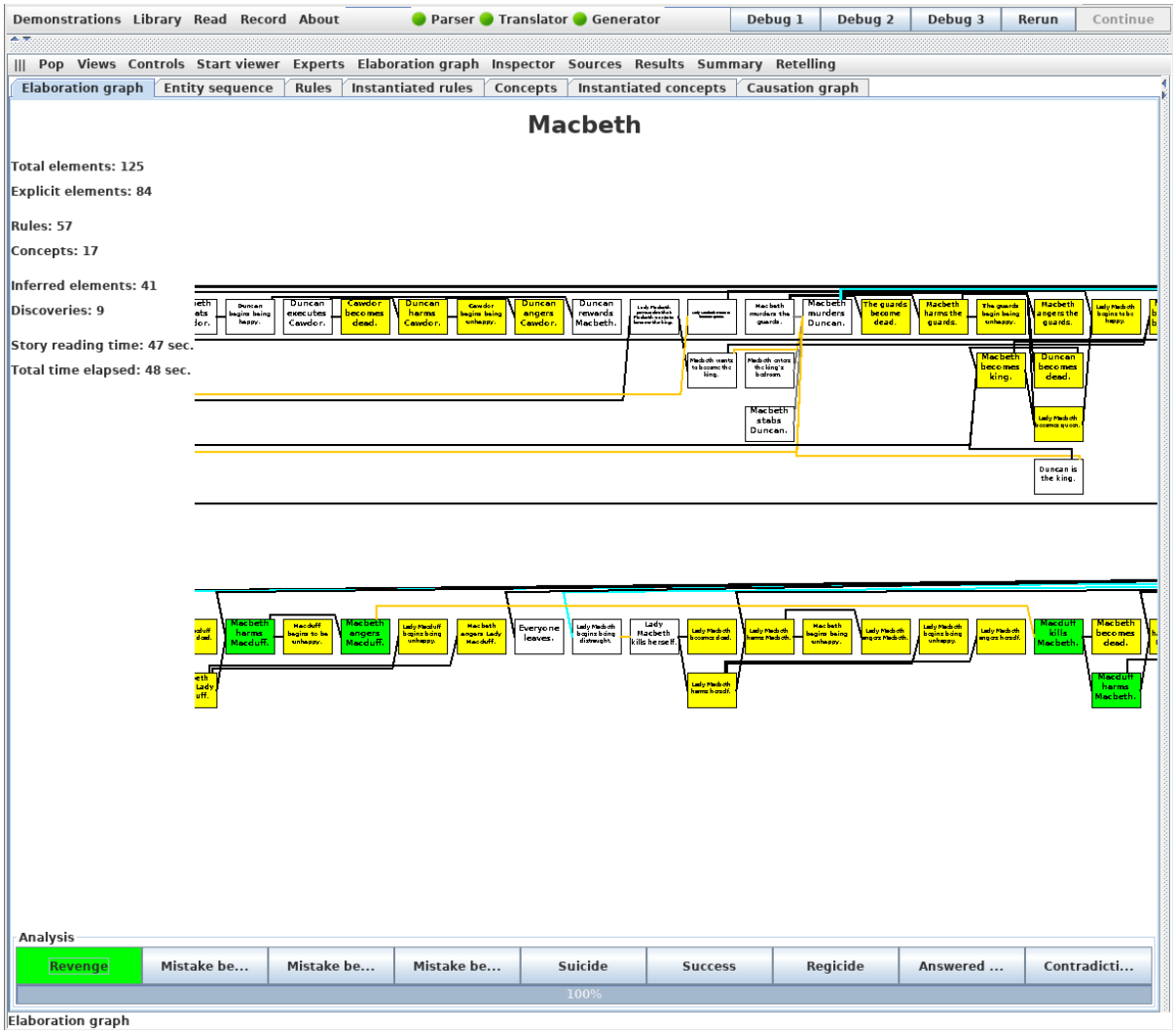


Figure 1: The *elaboration graph* shown here depicts the events in a simplified version of *Macbeth*, including deduced facts and conjectured causal connections. Concept patterns such as Revenge (highlighted in green) emerge from chains of such causes or inferences in the narrative.

a humanlike way, but are hardly ever expressed explicitly in the stories people encounter. Hence, we ourselves furnish Genesis with the necessary kind of background knowledge that even young children know. As a result of this general commonsense knowledge, provided through an auxiliary text file similarly

expressed in simple English, Genesis can discover many more causal and inferential connections than are expressed explicitly in the story, providing a richly connected graph.

There are two major structures for representing this commonsense information. The first is a family of different *commonsense rules*. Genesis possesses an arsenal of rule types each with a specialized behavior developed to meet a particular engineering need. Genesis possesses deduction rules, abduction rules, explanation rules, and unknowable-leads-to rules, among others. Instantiated rules are matched against the story, and when they fire, they may add new information to the story (such as deductive consequences or more detailed information about how an action could be carried out) or new connections (such as causal connections). These new story elements and connections are accumulated in the elaboration graph.

The second structure comprises *narrative concept patterns*. In the Genesis system, a concept pattern is a constellation of events in a story which together represent a high-level narrative theme such as *Success through adversity* or *Escalating violence* (See Lehnert (1981) for related work on plot units). Many, but not all, concept patterns involve *leads-to* relationships; that is, relationships that emerge from an unbroken chain of events and inferences in a story. For example, the concept pattern Revenge occurs whenever one act of harm is connected to a reciprocal act of harm through any number of intervening story elements (Figure 1).

B.1 Presumption rules fill in gaps

I have focused on a specific type of question: what would happen if we remove a particular element of the story. Enabling Genesis to simply remove an element and re-analyze the story is technically straightforward—most of the challenge there involves parsing the question, matching it against the story, and then re-running the story with Genesis's existing story understanding apparatus. The true technical challenge arises from the fact that a story with a missing element is not *simply* a shorter story. Consider, for example, the widespread ramifications of removals like these:

1. What if the assailant did not have a knife?
2. What if this character were not selfish?
3. What if the reader did not have a particular cultural background?
4. What if the sidekick had not left in the second act?

Each of these questions may drastically alter the outcome and interpretation of the story. And though these questions all have the same superficial form, they differ widely with respect to the kind of information they affect and the skills required to respond competently. All of the interesting things happen in the omission, so if you don't have the right latent background information—beyond the information you used to understand the original scenario—you will not be able to describe the alternative scenario intelligently!

What kind of background information is necessary, and how do we process it? For demanding what-if questions such as "What would happen if the sidekick had not left in the second act?", we might rely on an extensive and varied range of information, and the process for manipulating it might involve a lot of search and evaluation to find a plausible answer. For certain kinds of questions—which I call *gap-filling questions*—we seem to fill in gaps more or less automatically: we reflexively fill in missing information using commonsense knowledge and cognitive biases. Though aspects of this kind of unreflective filling-in can have unwelcome consequences—prejudice, functional fixedness (where we overlook new uses for familiar objects) (Duncker and Lees, 1945), hackneyed tropes, and jumping to conclusions, for example—it nevertheless lies at the core of our ability to make a story or a visual scene coherent by hallucinating missing details: using "the stereotypes of what we expected" (Minsky, 2006, Chapter 4), we can process a visual scene before we've seen every detail and we can read stories without needing every connection to be explicitly laid out.

To model this kind of reflexive gap-filling, I introduced *presumption rules* into the Genesis story understanding system. I extended Genesis's rule-matching system (which searches the story for elements that match the commonsense rules in its database, then adds inferences and causal connections to the story accordingly) to handle this new rule type and new behavior.

A presumption rule encodes fragile default knowledge about what to assume in the absence of evidence to the contrary. Hence, you can express many default assumptions as presumption rules such as the following:

- If someone enters the kitchen, then presumably that person wants to eat.
- If an adult stands at the front of a lecture hall, that person is presumably the instructor.
- There is smoke presumably because there is fire.

Declaring a presumption rule

The presumption rule type supplements Genesis's existing family of rule types in that presumption rules introduce genuinely *new, presumptive facts* into the story being read. This behavior is importantly different from the behavior of *deduction rules*, which only add completely certain conclusions ("If you kill someone, that person becomes dead.") and *explanation rules*, which can introduce new connections between existing events, but cannot introduce new events ("A character may kill someone because that character is angry.", interpreted as meaning "If both events occur in the story, tentatively add a causal connection between them").

The syntax for declaring presumption rules is consistent with the standard syntax for declaring other rule types. Presumption rules are signaled by the idiom "presumably" or "can" (which here means "could potentially"). The keywords "can" and "presumably" can be used interchangeably, and they can be used for rules formatted either as "if xx then yy" or as "yy because xx". For example, Genesis would recognize all of the following rules as presumption rules:

```
xx can enter the kitchen because xx wants to eat.

If xx stands in front of the lecture hall, then xx is
presumably the instructor.

If there is smoke, there can be fire.
```

As an aside, I note that our everyday language maintains subtly different rules for when *can* or *presumably* are the right word: in an inference rule, *can*

has a connotation of being "one presumption among many good alternatives", while *presumably* has a connotation of being "the one obvious presumption to make". For the purposes of this thesis, however, the two keywords can be used interchangeably.

Overshadowing

Because presumption rules encode *default* knowledge, they will only introduce a new event into the story if no other explanation exists. As such, presumption rules can become "overshadowed" by earlier rules that compete to provide an explanation. Presumption rules can be overshadowed by explicit sentences, or inferences, or explanation rules.

For example, consider the presumption rule "xx shoves yy presumably because xx dislikes yy". In a story, the following sentences would *match* this presumption rule because shoving occurs, but would preclude the rule from firing and introducing an explanation because in each case an explanation already exists:

- Riley shoves Casey because Riley and Casey are actors.
- Riley may shove Casey because Casey is in harm's way.

Conversely, presumption rules can introduce connections that overshadow explanation rules. Hence by controlling the order in which presumption rules and other rules fire, you can change which kind of explanation will dominate—the default presumption explanation or an alternate explanation. Such rule precedence provides a potential way to model differences in how attached people are to their presumptions. Some presumptions may be easily overridden; others, as in certain forms of psychopathology, are so firmly embedded that little can override them.

Future work for presumption rules

For the work in this thesis, the "presumptive" nature of presumption rules appears in two ways: first, the fact that they only fire to fill in explanatory gaps; second, the fact that they are intended to be used for abductive inference—inference

that is provisional and uncertain. This is the extent of the presumptive nature; once the presumption rules have fired, Genesis itself does not currently treat them any differently than other rule types. In particular, Genesis does not yet have the capacity to discard presumptions in light of new information.

In future work, however, I envision developing a more elaborate system for managing presumptions: not only introducing new presumptions, but comparing them against existing facts, choosing between competing presumptions, presuming large frameworks of knowledge rather than individual events, and revoking presumptions in light of new evidence. Such extensions would solidify the role of presumption rules as encoding fragile default knowledge.

Bibliography

Austin, J. L. (1956). A plea for excuses. In *Proceedings of the Aristotelian Society*, volume 57 of *New Series*. Blackwell Publishing.

Austin, J. L. (1971). A plea for excuses. In *Philosophy and linguistics*, pages 79–101. Springer.

Bentham, J. (1948). An introduction to the principles of morals and legislation (1789). *New York: Hafner*.

Boden, M. (1991). *The Creative Mind: Myths and Mechanisms*. Routledge.

Bradley, B. (2012). Doing away with harm. *Philosophy and Phenomenological Research*, 85(2):390–412.

Brandeis, L. and Warren, S. (1890). The right to privacy. *Harvard law review*, 4(5):193–220.

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. MIT Press.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT press, Cambridge, MA.

Davis, R. (1980). Meta-rules: Reasoning about control. *Artificial intelligence*, 15(3):179–222.

Dehghani, M., Tomai, E., Forbus, K. D., and Klenk, M. (2008). An integrated reasoning approach to moral decision-making. In *AAAI*, pages 1280–1286.

D'Ignazio, C. and Klein, L. F. (2020). *Data feminism*. MIT Press.

Doyle, J. (1980). *A Model for Deliberation, Action and Introspection.* PhD thesis.

Duncker, K. and Lees, L. S. (1945). On problem-solving. *Psychological monographs*, 58(5):i.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pages 160–171. PMLR.

Forbus, K. D. (1984a). Qualitative process theory. *Artificial Intelligence*, 24:85–168.

Forbus, K. D. (1984b). Qualitative process theory. *Artificial intelligence*, 24(1-3):85–168.

Gerstenberg, T. and Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, 216:104842.

Gerstenberg, T., Zhou, L., Smith, K. A., and Tenenbaum, J. B. (2017). Faulty towers: A hypothetical simulation model of physical support. In *CogSci*.

Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition*. Psychology Press.

Goldstein, J. S. (1992). A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution*, 36(2).

Graeber, D. (2012). *Debt: The first 5000 years*. Penguin UK.

Grice, H. P. (1975). Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Hanna, N. (2016). Harm: omission, preemption, freedom. *Philosophy and Phenomenological Research*, 93(2).

Hayes, P. J. (1979). The naive physics manifesto. *Expert systems in the microelectronic age*.

Holmes, D. A. (2017). *Story-enabled hypothetical reasoning*. PhD thesis, Massachusetts Institute of Technology.

Jackson, N. (2016). Moral particularism and the role of imaginary cases: A pragmatist approach. *European Journal of Pragmatism and American Philosophy*, 8(VIII-1).

Kamm, F. M. (2008). *Intricate ethics: rights, responsibilities, and permissable harm*. Oxford University Press.

Kamm, F. M. (2011). *Ethics for Enemies: Terror, Torture, and War*. Oxford University Press.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. MIT Press, Cambridge, MA.

Langley, P. (2012). The cognitive systems paradigm. *Advances in Cognitive Systems*, 1(1):3–13.

Lefkowitz, D. (2008). On the concept of a morally relevant harm. *Utilitas*, 20(4):409–423.

Lehnert, W. G. (1981). Plot units and narrative summarization. *Cognitive Science*, 5(4):293–331.

Magid, R. W. and Schulz, L. E. (2017). Moral alchemy: How love changes norms. *Cognition*, 167:135–150.

Magid, R. W., Sheskin, M., and Schulz, L. E. (2015). Imagination and the generation of new ideas. *Cognitive Development*, 34:99–110.

Marr, D. (1982). *Vision*. W.H. Freeman, San Francisco, CA.

McCarthy, J. (1980). Circumscription—a form of non-monotonic reasoning. *Artificial intelligence*, 13(1-2):27–39.

McCarthy, J. et al. (1960). *Programs with common sense*. RLE and MIT computation center.

Mill, J. S. (1859). Utilitarianism (1863). *Utilitarianism, Liberty, Representative Government*.

Minsky, M. (1994). Negative expertise. *International Journal of Expert Systems*, 7(1):13–19.

Minsky, M. L. (2006). *The Emotion Machine*. Simon and Schuster, New York, NY.

Nagel, T. (1979). Fragmentation of values. In *Mortal Questions*.

Nagel, T. et al. (1979). Moral luck.

Newell, A., Shaw, J. C., and Simon, H. A. (1959). Report on a general problem solving program. In *IFIP congress*, volume 256, page 64. Pittsburgh, PA.

Nozick, R. (1974). *Anarchy, state, and utopia*, volume 5038. New York: Basic Books.

Nussbaum, M. C. (2006). *Frontiers of justice: Disability, nationality, species membership*. Belknap Press Cambridge, MA.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. In *Images, perception, and knowledge*, pages 1–36. Springer.

Radul, A. (2009). *Propagation networks: A flexible and expressive substrate for computation*. PhD thesis, Electrical Engineering and Computer Science Department, MIT, Cambridge, MA.

Rissland, E. L., Ashley, K. D., and Branting, L. K. (2005). Case-based reasoning and law. *Knowledge Engineering Review*, 20(3):293–298.

Ryff, C. D. and Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of personality and social psychology*, 69(4):719.

Saeed, J. I. (2015). *Semantics: Introducing Linguistics*. Wiley-Blackwell.

Saxe, R. (2016). Moral status of accidents. In *Proceedings of the National Academy (PNAS)*, volume 113, pages 4555–4557.

Schulman, S. (2016). *Conflict is not abuse: Overstating harm, community responsibility, and the duty of repair*. arsenal pulp press.

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological review*, 74(1):29.

Singh, P. (2005). *EM-ONE: an architecture for reflective commonsense thinking*. PhD thesis, Massachusetts Institute of Technology.

Sloman, A. (2015a). Impossible objects. URL: http://www.cs.bham.ac.uk/research/projects/cogaff/misc/impossible.html. Accessed: 2017-02-23.

Sloman, A. (2015b). Some (possibly) new considerations regarding impossible objects.

Speer, R., Havasi, C., and Lieberman, H. (2008). Analogyspace: Reducing the dimensionality of common sense knowledge. In *Aaai*, volume 8, pages 548–553.

Spelke, E. S. and Kinzler, K. D. (2007). Core knowledge. *Developmental science*, 10(1):89–96.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 49:433–

460.

Von Wright, G. H. (1951). Deontic logic. *Mind*, 60(237):1–15.

Williams, B. M. (2017). *A commonsense approach to story understanding*. PhD thesis, Massachusetts Institute of Technology.

Williams, B. M., Lieberman, H. A., and Winston, P. H. (2017). Understanding stories with large-scale common sense.

Winograd, T. (1971). Procedures as a representation for data in a computer program for understanding natural language. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC.

Winston, P. H. (1970). *Learning structural descriptions from examples*. PhD thesis, Electrical Engineering and Computer Science Department, MIT, Cambridge, MA.

Winston, P. H. (2011). The strong story hypothesis and the directed perception hypothesis. *AAAI*.

Winston, P. H. (2012a). The next 50 years: a personal view. *Biologically Inspired Cognitive Architectures*.

Winston, P. H. (2012b). The right way. *Cognitive Systems Foundation*.

Winston, P. H. (2018). Self-aware problem solving. Technical report.

Winston, P. H. and Holmes, D. (2018). The Genesis enterprise: Taking artificial intelligence to another level via a computational account of human story understanding. Technical report.

Wittgenstein, L. (1979). Remarks on frazer's golden bough. *The MyThology in our*, page 29.

Woods, W. A. (1975). What's in a link: Foundations for semantic networks. In *Representation and understanding*, pages 35–82. Elsevier.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT press.

Young, L., Cushman, F., Hauser, M., and Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of*

the National Academy of Sciences, 104(20):8235–8240.

Zadeh, L. A., Klir, G. J., and Yuan, B. (1996). *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers*, volume 6. World Scientific.