Advances in Cognitive Systems 4 (2016) Submitted 4/2016; published 6/2016

Story-enabled hypothetical reasoning

Dylan Holmes DXH@MIT.EDU
Patrick Winston PHW@MIT.EDU
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139 USA

Abstract

Varieties of hypothetical reasoning underlie much of intelligent behavior, from detecting low-level visual affordances about what surfaces are graspable, to arguing about laws in terms of counterfactuals. We have built three illustrative programs which reason about hypothetical circumstances by deploying mechanisms for filling in gaps during story comprehension. The first program evaluates culpability based on what could have happened but didn't in a simple breaking-and-entering scenario. The second program re-evaluates a scenario based on the Russian-Estonian cyberwar of 2007 in terms of the participants' differing outlooks. The third program judges a child based on the alternative actions he could have taken to get control of another child's ball, but didn't. Each program is built upon the Genesis story-understanding system.

From a research perspective, we consider hypothetical reasoning—the ability to conceive and cogently discuss multiple *possibilities*—to be an essential aspect of our overall attempt to understand how human-level intelligence is different from that of non-human animals. From an engineering perspective, we believe that to behave masterfully in the present, systems must be conversant in theories of their own operation and the trajectory of the world. In this paper, we focus on hypothetical reasoning of a kind enabled by a substrate developed for telling and understanding stories. Our goal is to lay a foundation for future systems with capabilities that will be increasingly important in proportion to their power and deployment.

- Systems that realistically predict how they would operate under various extreme circumstances.
- Systems that are able to anticipate hazards before they happen, selecting relevant futures out of a multitude of possible ones.
- Systems that are able to construct plausible explanations for what could have led to the present circumstances.
- Systems that justify their decisions in light of alternatives they are trying to preclude.
- Systems that deploy knowledge of precedent when predicting what will happen next, which actions will lead to favorable outcomes, and which plans are likely dead ends.

Elaboration graph | Entitysequence | Rules | Instantiated rules | Concepts | Instantiated concepts

Total elements: 0

Explicit elements: 0

Rules: 2

Concepts: 3

Inferred elements: 0

Discoveries: 0

Story reading time: 0 sec.

Total time elapsed: 0 sec.

Analysis

100

Mental Models

s Summary Retelling

cepts

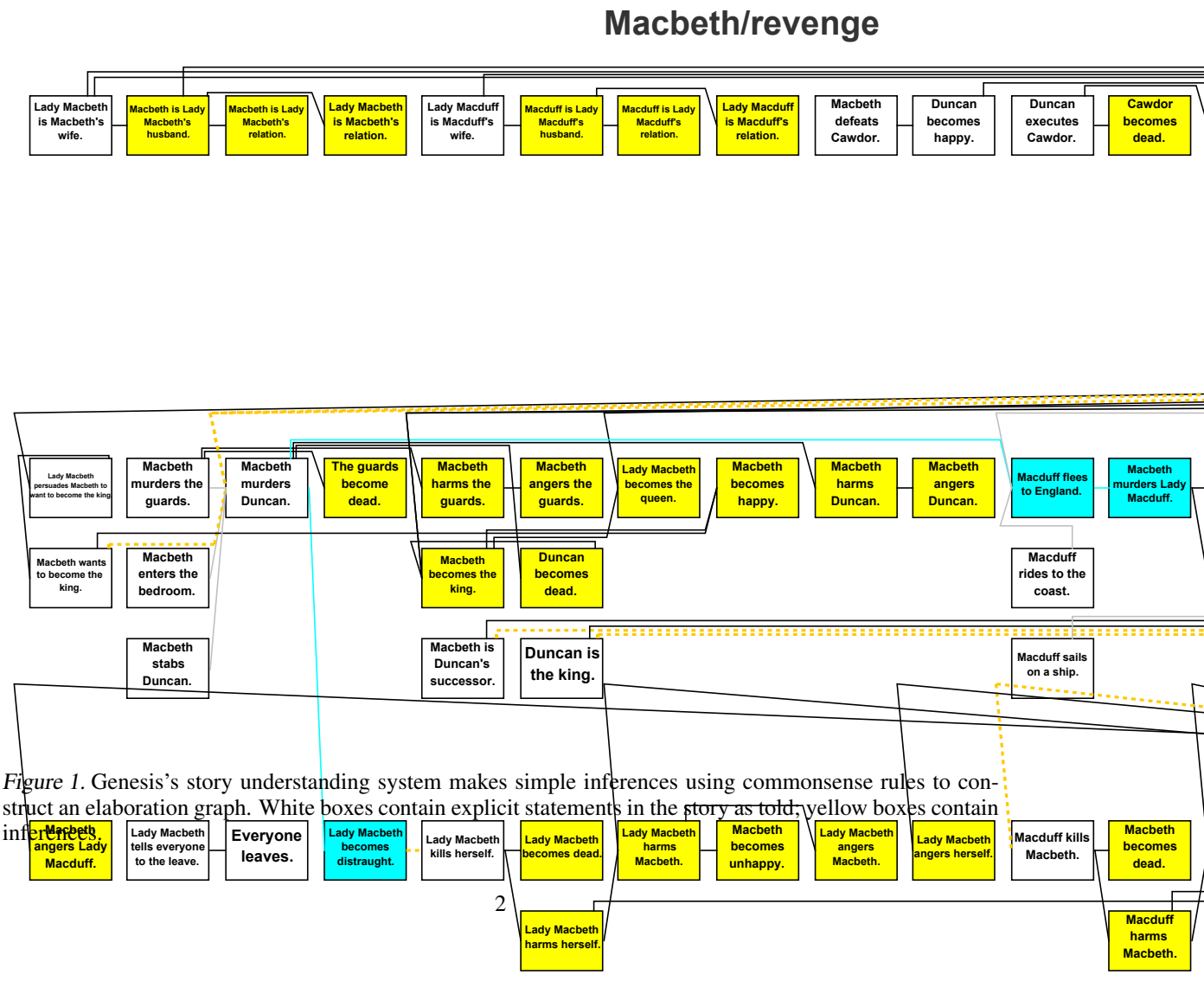

Figure 1. Genesis's story understanding system makes simple inferences using commonsense rules to construct an elaboration graph. White boxes contain explicit statements in the story as told; yellow boxes contain inferences.

Pyrrhic victory | Mistake because harmed | Mistake because harmed | Mistake because unhappy

Figure 2. Genesis discovers an instance of Revenge by searching the elaboration graph for the Revenge concept pattern.

Like the commonsense rules, concept pattern descriptions are provided to the Genesis system in plain English, generally with leads-to relations (i.e. relations between events that are connected through any number of intermediate events). Here are two examples of *revenge*:

- Revenge 1: X's harming Y leads to Y's harming X;
- Revenge 2: X's harming Y leads to Y's wanting to harm X. Y's wanting to harm X leads to Y's harming X.

The right version depends on the thinker, so we are able to model specific thinkers by including more or less sophisticated or more or less biased ways of looking at the world.

Note that, according to the connections in the graph fragment in Figure 2, Macduff killed Macbeth because Macbeth angered Macduff. Fortunately, we do not always kill the people who anger us, but in the story, as given, there is no other explanation, so Genesis inserts the connection using an explanation rule, believing the connection to be plausible in the absence of any other reason.

2. Justification based on what could have happened

Having provided an introduction to the Genesis story-understanding substrate, we now turn to the first of the three hypothetical reasoning programs which we have developed.

One of the fundamental skills required for robust reasoning is the ability to explain decisions in terms of what could possibly have happened—but didn't. Consider the ability to assess a plan not just in terms of whether it works, but in terms of the contingencies it accounts for. Consider the ability to reason about legal justifications (such as self-defense) that depend on threats which are intercepted before they can happen (Rissland [1989]). Or consider the ability to direct such hypothetical ability inward, to produce self-monitoring machines that can elaborate on their own reasons for acting (Forbus and Hinrichs [2006]).

We have constructed a rudimentary demonstration that models this sort of ability to provide explanations in terms of hypotheticals. By using Genesis's ability to fill in explanatory gaps and to compare story outcomes against precedent, our punctured-stories program can propose reasonable answers to questions of the form "What would happen if . . . ?" Here is an example scenario:

> Alex and Martha have despised each other for a long time. The hour is late; George and Martha are asleep. Martha wakes up because Alex breaks a window. Alex begins shouting, then Alex brandishes a knife. Martha shoots Alex; Alex dies.

From a story-understanding view, what often builds tension and drama and rising action in a story is what threatens to happen—but hasn't yet. Moreover, the memorability of a story is often

cached out in terms of surprise—what was supposed to happen, but didn't. In order to have a human-like sense of pacing of a story, Genesis must be able to suppose what will happen—then become apprehensive or surprised, accordingly. In this scenario, Genesis identifies the knife as a source of potential—not actual—harm, and uses it to find the concept pattern of *Martha's self-defense* in the story.

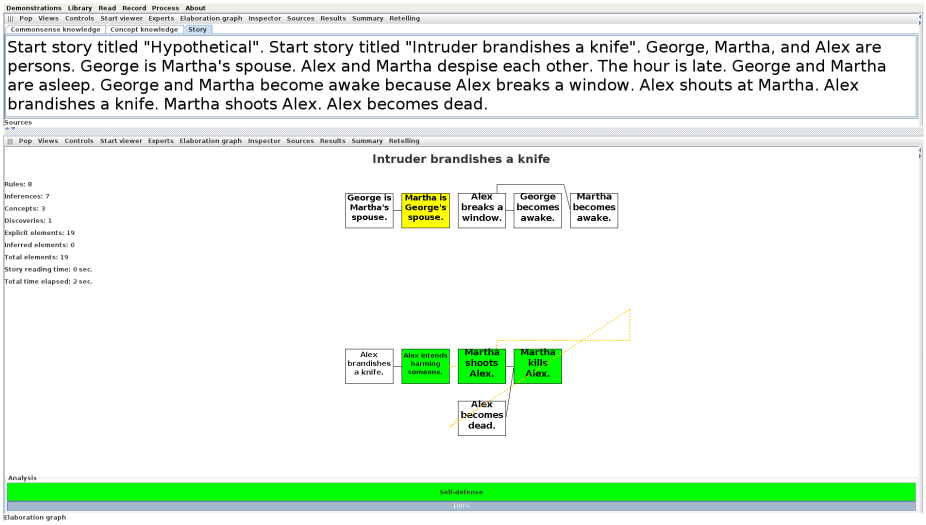

Figure 3. The initial setup of the story. Through a sequence of tentative (explanation-rule) and deductive (deduction rule) inferences, Genesis detects the *self-defense* concept pattern (highlighted in green.)

We can query the system's understanding by asking, in English:

"What would happen if Alex didn't brandish a knife?"

In response, Genesis removes the event "Alex brandishes a knife" from the story and re-analyzes it to see the result of this difference. Immediately, it concludes that "self defense" no longer fits as an explanation of Martha's actions and searches for another motive in order to make sense of the story. Linking putatively related events, the program connects the shooting with a presumption that "Martha despises Alex," and decides that Martha's action is unjustified, labeling it *Martha's spiteful violence*.

To enable this capability, we developed a new type of rule which we call a *presumption rule*; like *explanation rules*, discussed in the Genesis overview, presumption rules tentatively fill in explanatory gaps in the story. But where explanation rules tentatively introduce only causal connections, presumption rules may also introduce new putative *facts* into the story.

Here is a play-by-play explanation of how Genesis processes this scenario:

- Initially, the events "Martha shoots Alex" and "Alex brandishes a knife" are explicitly mentioned in the story.

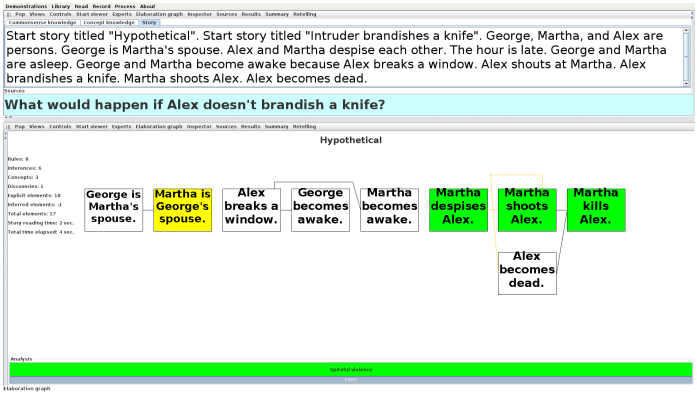

Figure 4. When Genesis analyzes a hypothetical alternative story, without a perceived threat of harm, Genesis discovers an alternative explanation, *spiteful violence*.

- Genesis infers that Alex brandishes a knife *presumably because* Alex intends to harm someone. Thus, through this presumption rule, Genesis introduces "Alex intends to harm someone" as well as its putative connection to Martha's shooting.
- An alternate explanation — that Martha *may* shoot Alex because Martha despises Alex — is available, but unused. Because the explanation "Alex intends to harm someone" already exists, this may-rule does not fire.

This outcome is sensitive to the ordering of the rules; indeed, different rules, concept patterns, and rule orderings could be used to model different reader perspectives and allow for different presumptive behavior in different settings. For example, models of common sense reasoning might use many unmonitored presumptive inferences, whereas models of analytical reasoning might rely more on conservative deductive inference and less on presumptive inference.

The analysis in this scenario shows how hypothetical reasoning enables a kind of simple reflective thinking about possibilities and differences. In the next stage of processing, a high-level supervisory system takes this capability to another level by comparing the two stories in memory and analyzing how its own reasoning processes differed between the two stories. Here is its own report, automatically compiled and generated in English from its own trace of its work:

From an event-based perspective, I note the following changes:

- It's no longer the case that "Alex intends to harm someone because Alex brandishes knife".
- It's no longer the case that "Martha shoots Alex, probably because Alex intends harming someone."

From a thematic perspective, the following concepts disappear:

- Alex's self-defense

. . . and the following concepts are introduced.

- Martha's spiteful violence

We see this kind of analysis as marking the first steps toward a self-reflective, iterated story analysis: by considering many alternative possibilities and other potential outcomes, Genesis enriches its understanding of the main plot in a human-like way. The ability to infer potential harm serves as a precursor to ethical deliberation and to a reader's sense of suspense. In future work, we aim to extend this facility with possibility so as to model high-level cognitive phenomena such as moral development and case-based legal reasoning, as well as reader reactions such as suspense and surprise.

3. Political decisions

When performing moral decision-making in the George/Martha scenario, the program concluded that Martha's act was of self-defense because a knife constituted a threat—a *potential* harm. Evidently, stories about potential outcomes are important when reasoning about the seriousness of a moral transgression. One missing feature of this reasoning was an explicit moral evaluation: a conclusion about the wrongness and seriousness of each act.

In this section, we illustrate moral evaluation with an application of hypothetical reasoning in international policy. International policy is a fruitful area for studying moral evaluation, because moral evaluation is defined in terms of one's value system, and international politics involves empathizing with and negotiating between different value systems. If we can model the value systems that guide nations and individuals, we can begin to model the way in which people with different cultural contexts and background knowledge think. Here, we focus on a particular rudimentary case of cultural difference, namely differences in *allegiance*.

The scenario we have chosen for analysis is the Russia-Estonia cyberwar of 2007, long a part of Genesis's story repertoire.

Here, the original story presumes an Estonian background; this allegiance has many ramifications at the rule level, which lead to the emergence of nationalistic concept patterns.

We can ask, however, how the picture would be different if the reader were an ally of Russia rather than Estonia. We ask, in English, "What would happen if I am not from Estonia?". The system removes the relevant facts from the story, and attempts to predict the meaningful differences (Figure 6).

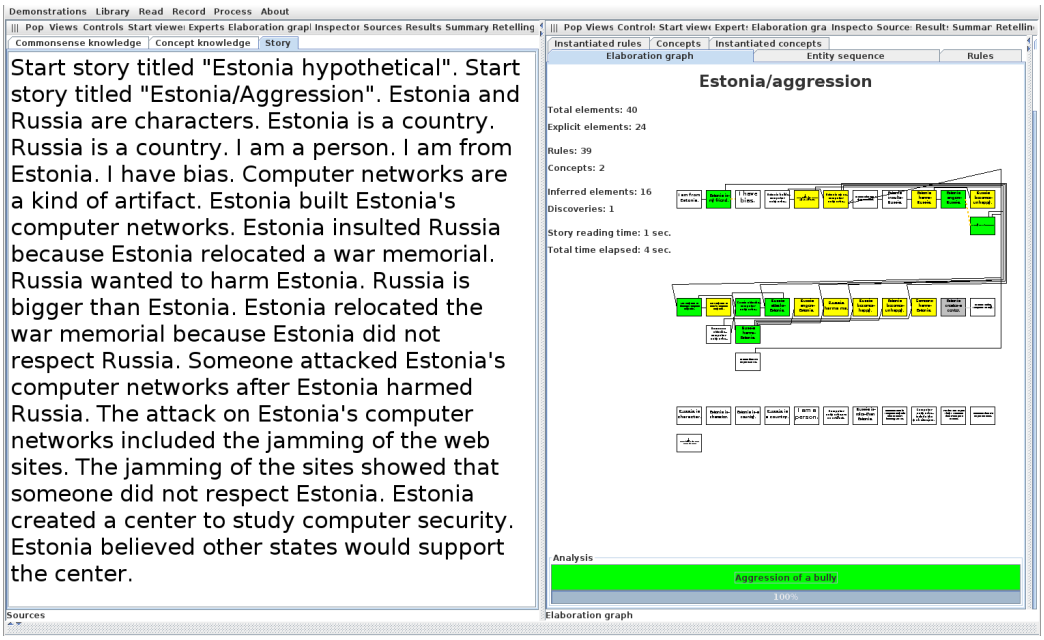

Figure 5. In the initial setup, Genesis reads about the Estonia-Russia conflict while taking an Estonia-sympathetic perspective. The left pane shows the text of the story, which is essentially a neutral reporting of events. The right pane shows the resulting elaboration graph structure, including the concept pattern (highlighted in green) for *aggression of a bully* [Russia].

As a result of the difference in perspective, the situation changes from being the Aggression of a Bully (cached out as "When my friend angers someone and that person retaliates.") to Teaching a Lesson (the dual concept, cached out as "When someone angers my friend and my friend retaliates.")

This program illustrates a key capability, namely being able to interpret a story from many different listeners' perspectives. From a scientific perspective, we envision being able to extend this program to model theory of mind and empathy, as well as cultural differences in background assumptions, sacred values, literary allusions, and common-sense knowledgeMinsky. From an engineering perspective, we envision developing this program into a tool which is as useful to a political analyst or diplomat as spreadsheets are to a financial analyst—allowing people to explore the effects of political actions for various demographic groups and in various scenarios.

4. Trait inference based on what didn't happen

To be robust, systems must know much more than the strategies that they actually end up using; they must know at each point something about what they would do otherwise, and what recourse they have if something goes wrong. In our third program, we illustrate how humans pass moral judgment and assign character traits based on actions characters could have hypothetically taken. For example, in everyday life, we may consider a person vicious if that person chooses violent means to achieve their ends when we know that there are other effective alternatives. In this way, we move from

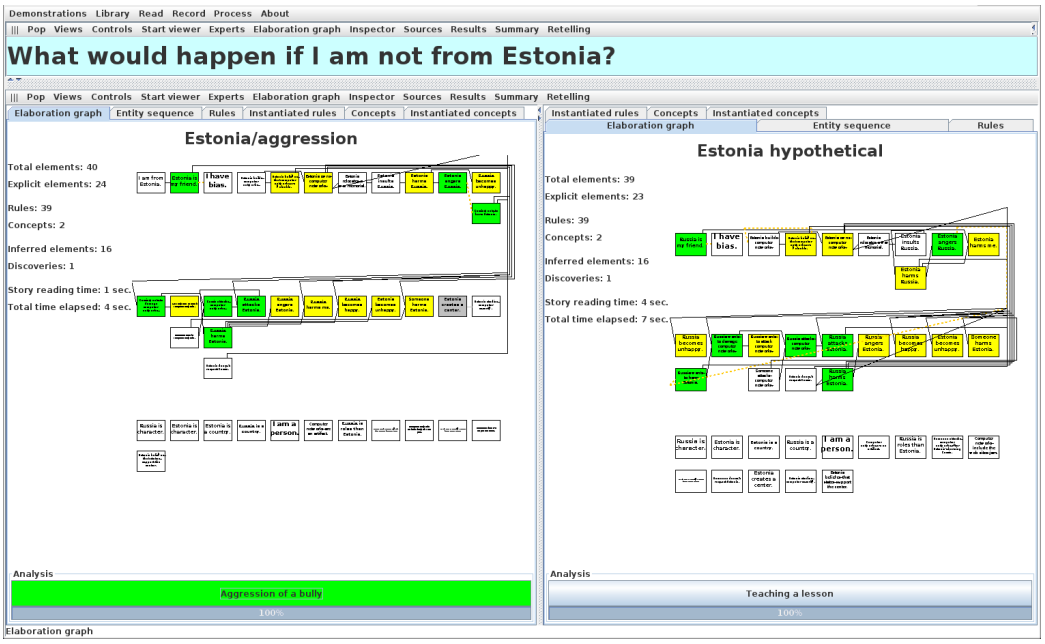

Figure 6. Genesis perceives significantly different narrative structure when adopting a hypothetical Russia-sympathetic perspective. The view on the left shows the original setup as it appeared in Figure 5, where *aggression of a bully* is the dominant concept pattern. The view on the right shows the rendering of the story where Genesis assumes a Russia-sympathetic view. Here, the *teaching a lesson* concept pattern dominates instead.

the allegiance-based model of morality in the Estonia/Russia example, to a playground model of morality based on the actions characters didn't take.

In detail, the program is presented with a story, then asked to infer the goals of each character and to make moral judgments about the means by which characters are achieving their goals. The program possesses a database which relates possible goals to different means of achieving those goals, as well as the consequences of various means. Specifically, the program possesses a list of hand-coded sentences such as "In order for `xx` to have `ww`, `xx` can take `ww` from `yy`". The program first infers what characters want (by aligning their explicit actions with means such as "`xx` can take `ww` from `yy`"), then searches for other solutions to their putative goals (by finding other methods for achieving the same goal, such as "In order for `xx`= to have `ww`, `xx` can *ask* `yy` for `ww`.")

In one scenario, the program is presented with a story involving playground interactions between two characters. The first character wants a ball that the second character is currently using—and so the first character simply takes it. We ask the program to evaluate the scenario under two conditions: in the first condition, the program knows about *several* means of obtaining the ball—for example, taking it away (as happened in the story) or requesting it. In the second condition, the program knows only about obtaining a ball by taking it away. (As a sort of control instance, we also show what happens when the character chooses to *ask* for the ball rather than take it.)

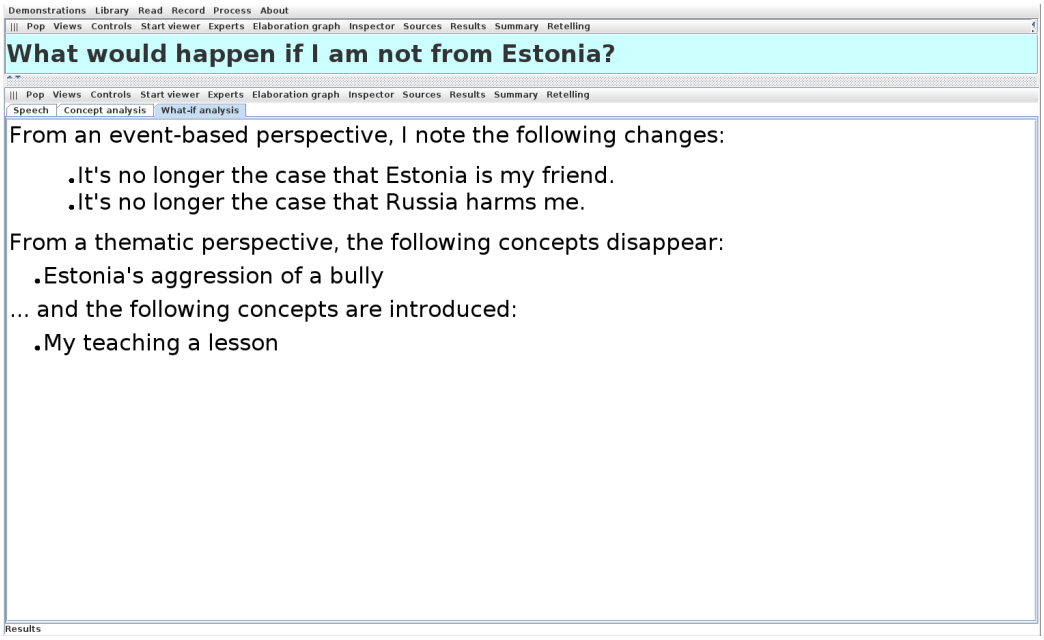

Figure 7. As in the previous section, Genesis can analyze and distill the differences between the two versions of the story at several levels of granularity—event-based and theme-based. Genesis automatically generates the English in each bullet point from its internal representation of the rules and sentences in the story.

In both conditions, the program correctly infers the first character's goal by aligning what happened in the story ("taking the ball") with a related goal in the database, represented as a template story ("If `x` has `y` and `z` wants `y`, `z` may take `y` from `x`".) Also in both stories, the program acknowledges the negative consequences of such an action ("If `x` has `y` and `z` takes `y` from `x`, then `x` presumably becomes sad.")

The difference is that in the case where the program has access to alternatives, it reports that Patrick is brutish; Patrick could have acquired the ball by asking for it—but chose instead to take it away. In the second case, with limited knowledge of alternatives, it reports "I believe that Patrick behaved wrongly—but I know of no other way the character could have achieved this goal." From the program's point of view, the act was a kind of unavoidable cost of achieving the goal.

In this way, our program models several interesting aspects of moral reasoning, with implications for models of children's budding moral reasoning ability. First, it reasons about goals and motives in a feedback loop where character actions imply certain means and ends, and those actions are subsequently judged based on alternative means. Second, it assigns character traits based on what viable alternatives were available. Such hypothetical reasoning was made possible by a library of commonsense information deployed in the form of stories—and although our example is merely an illustration, we believe it captures the general point that robust reasoning—specifically about hypothetical situations—is inextricable from sophisticated world-knowledge and the ability to align stories with precedents. Third, it constitutes a simple model of child morality. The system is "childlike" because it does not yet possess the reflective capability to question goals themselves

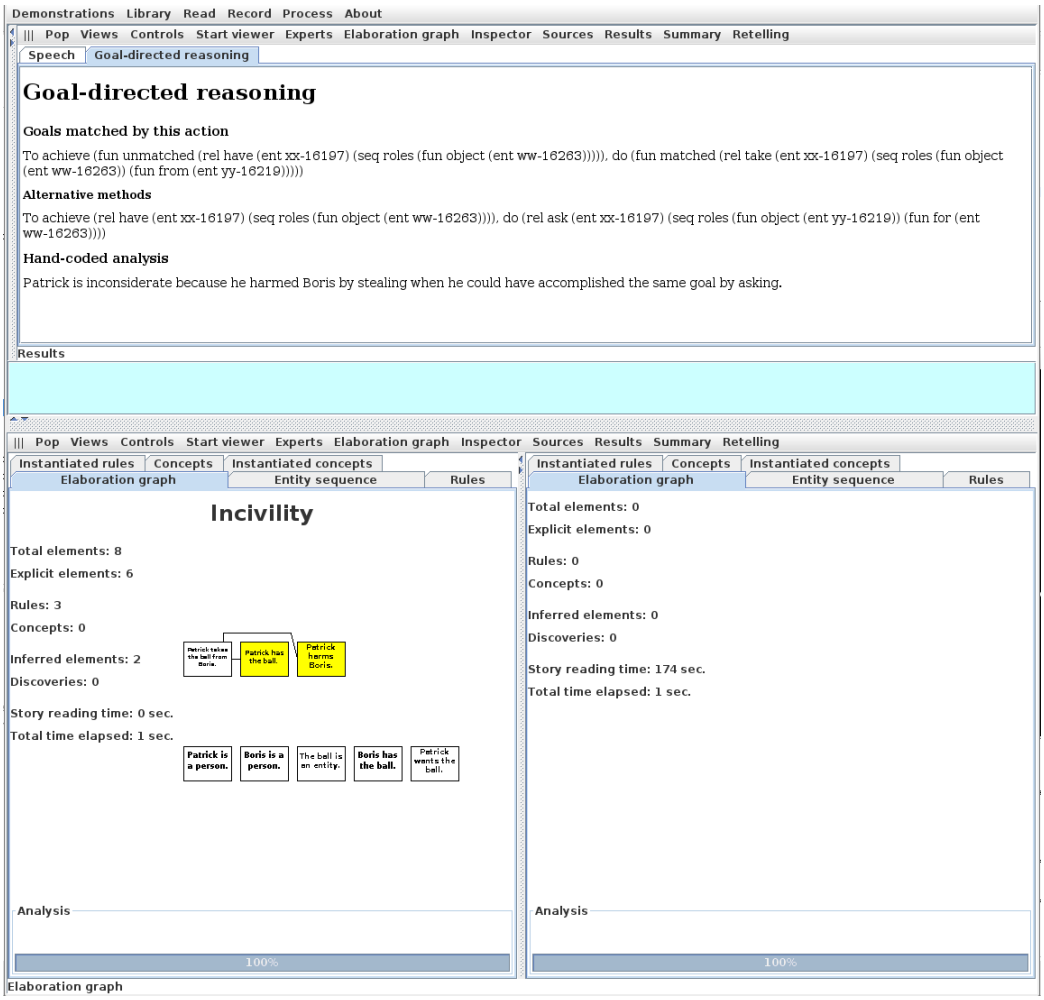

Figure 8. In the first setup, the evaluating program is given knowledge of several goals, methods for achieving them, and side-effects of each method. Hence the program concludes that Patrick is inconsiderate for taking the ball when asking might have caused less harm.

or to perform sophisticated cost-benefit analysis. Moreover, it mimics the crude and insensitive behavior of children who are still learning about prosocial ways of achieving their goals. In human adults, such methods may be replaced, on reflection, with more diplomatic means. In this system, additional knowledge provides additional possibility—and additional responsibility. (As the number of viable alternatives increases, the characters' culpability in choosing one of the unethical alternatives grows.) Fourth and finally, our program highlights the exculpatory nature of extreme circumstances: when the knowledge base of the system is artificially limited, it produces the same kind of excuse ("it was wrong, but unavoidable") that we often use to describe choices we make in dire circumstances.

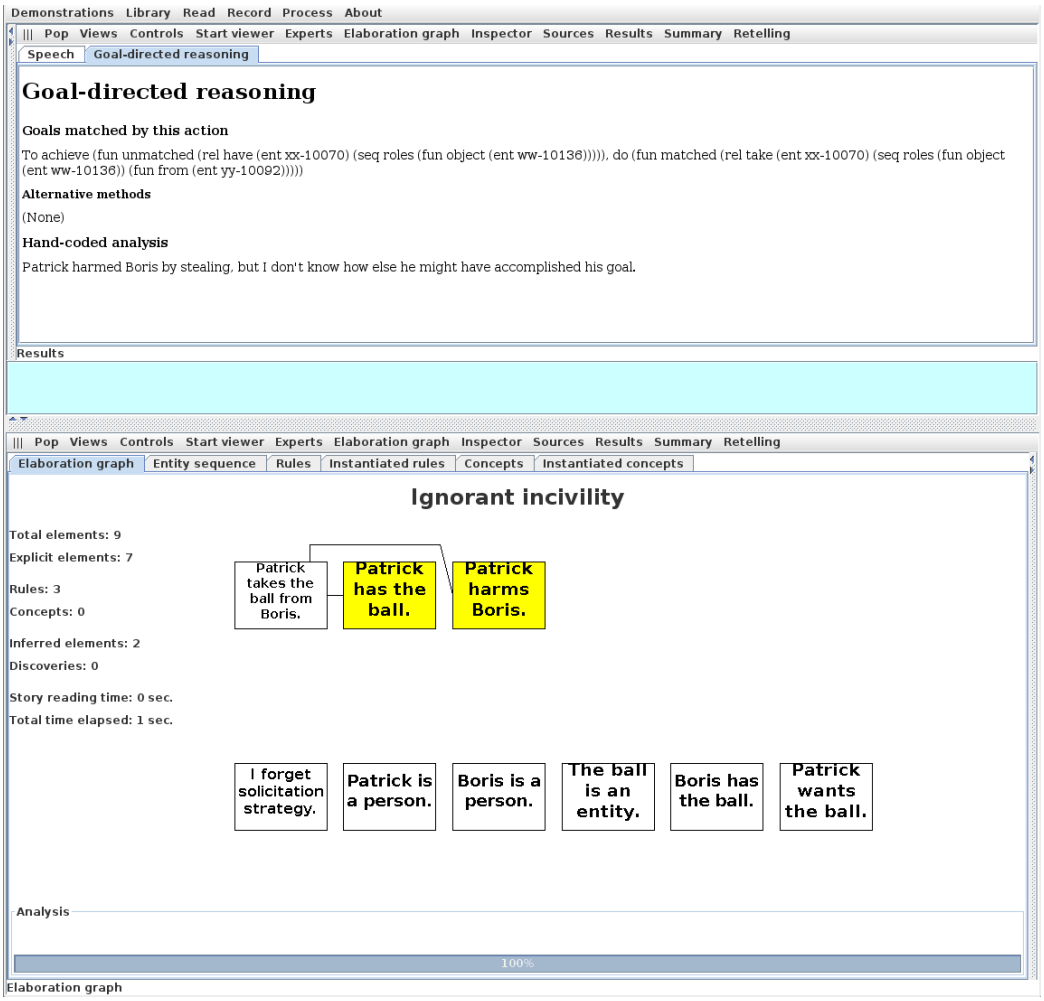

Figure 9. Without knowledge of multiple means, the program becomes confounded: Patrick harmed Boris by taking the ball, but the program knows of no other method for achieving the same goal. This sort of ambivalent reasoning imitates how humans seem to make judgments about morally exigent circumstances.

5. What's next

We have just begun to explore the possibilities of using hypothetical reasoning to infer means, ends, and character traits. There is much more to do in terms of richer representations of goal-stories, such as tradeoffs involved in one method over another, and what each goal can achieve and what their side-effects are. We can produce a more sophisticated ("more mature") reasoning ability by taking into account particular extenuating circumstances, accidents (which mitigate culpability), and intentionality (which magnifies culpability).

We believe that hypothetical reasoning is the right setting for many different sorts of cognitive behaviors; we list a small sample of them in the table shown in Figure 10.

Field of analysis	How would the analysis change if . . .
Case-based reasoning in medicine	. . .the patient's T-cell count were diminished?
Case-based reasoning in morality and law	. . .the suspect did not have a weapon?
Social psychology	. . .I look for situational explanations, rather than trait-based explanations?
Conflict resolution, empathy, diplomacy	. . .I read the story with this particular cultural outlook?
Moral development, self-modeling, child psychology	. . .I steal this toy when no one is looking?
Story trope analysis, personality traits, story-generation	. . .Red Riding Hood were the villain?
Literary analysis, reasoning from precedent, analogical alignment	. . .I compare this novel to *The Great Gatsby*?
Planning, naïve physics, on-the-fly safety analysis	. . .I run down the street with a full bucket of water?

Figure 10. Varieties of hypothetical reasoning enable many different cognitive capabilities.

6. Contributions

In this paper, we describe three programs we have developed on top of the Genesis story-understanding system. Each program applies story-based hypothetical reasoning to answer a different type of moral question. The first evaluates hypothetical variations in a story, making moral evaluations based on what could have happened, but didn't. The second program evaluates hypothetical variations in a reader/listener, making situational judgments from a perspective it does not itself occupy. The third evaluates hypothetical variations in goal-directed strategy, making character judgments based on the actions characters could have taken—but didn't.

To implement these programs, we developed a suite of new tools and capabilities on top of the existing Genesis substrate: first, we introduced presumption rules, which capture the frame-like default assumptions that people routinely make when filling in gaps in a story. Next, we demonstrated how collections of presumption rules can enable Genesis to answer hypothetical questions on a variety of subjects—moral counterfactual questions about self-defense (in the first program), and diplomatic questions about different audiences (in the second program). Finally, in the third program, we showed how hypothetical reasoning is the right way to model certain character traits

and judgments, as summarized by the slogan: "Your character is often defined by what you *could* have done—but didn't."

References

Matthew Paul Fay. Enabling imagination through story alignment. Master's thesis, Electrical Engineering and Computer Science Department, MIT, Cambridge, MA, 2012.

Kenneth D. Forbus and Thomas R. Hinrichs. Companion cognitive systems: a step toward human-level ai. *AI Magazine*, 2006.

Caryn Krakauer. Story retrieval and comparison using concept patterns. Master's thesis, Electrical Engineering and Computer Science Department, MIT, Cambridge, MA, 2012.

Marvin Lee Minsky. *The Emotion Machine*. Simon & Schuster.

Michael W. Morris and Kaiping Peng. Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality and Social Psychology*, 67(6):949–971, 1994.

E. L. Rissland. Dimension-based analysis of hypotheticals from supreme court oral argument. In *Proceedings of the 2Nd International Conference on Artificial Intelligence and Law*, ICAIL '89, pages 111–120, New York, NY, USA, 1989. ACM. ISBN 0-89791-322-1. doi: 10.1145/74014.74030. URL http://doi.acm.org/10.1145/74014.74030.

Sila Sayan. Audience aware computational discourse generation for instruction and persuasion. Master's thesis, Electrical Engineering and Computer Science Department, MIT, Cambridge, MA, 2014.

Patrick Henry Winston. The strong story hypothesis and the directed perception hypothesis. *AAAI*, 2011.

Patrick Henry Winston. The right way. *Cognitive Systems Foundation*, 2012a.

Patrick Henry Winston. The next 50 years: a personal view. *Biologically Inspired Cognitive Architectures*, 2012b.

Patrick Henry Winston. Model-based story summary. In Mark A Finlayson, Ben Miller, and Remi Ronfard, editors, *Proceedings of the 6th Workshop on Computational Models of Narrative(CMN 2015)*, volume 45. Oasics, 2015.