The Curious Case of Commonsense Intelligence

Yejin Choi

Commonsense intelligence is a long-standing puzzle in AI. Despite considerable advances in deep learning, AI continues to be narrow and brittle due to its lack of common sense. Why is common sense so trivial for humans but so hard for machines? In this essay, I map the twists and turns in recent research adventures toward commonsense AI. As we will see, the latest advances on common sense are riddled with new, potentially counterintuitive perspectives and questions. In particular, I discuss the significance of language for modeling intuitive reasoning, the fundamental limitations of logic formalisms despite their intellectual appeal, the case for on-the-fly generative reasoning through language, the continuum between knowledge and reasoning, and the blend between symbolic and neural knowledge representations.

Commonsense intelligence is a long-standing challenge in AI. Despite considerable advances in deep learning, AI systems continue to be narrow and brittle. One of the fundamental limitations of AI can be characterized as its lack of commonsense intelligence: the ability to reason intuitively about everyday situations and events, which requires rich background knowledge about how the physical and social world works.[1]

Trivial for humans, acquiring commonsense intelligence has been considered a nearly impossible goal in AI. In fact, until several years ago, the word "commonsense" was considered taboo for anyone wanting to be taken seriously in the mainstream research community. How, then, is this goal *now* feasible? To help answer this question, we will characterize what approaches have been tried in the past and what alternative paths have yet to be explored.

First and foremost, the significance of language – not just words and phrases, but the full scope of natural language – has long been overlooked as a representation medium for modeling commonsense knowledge and reasoning. At first glance, language seems too imprecise and variable, thus, many earlier efforts sought logic-based formalisms to describe commonsense rules for machines. But despite their intellectual appeal, logic-based formalisms proved too brittle to scale beyond experimental toy problems. In contrast, *language-based formalisms,* despite their apparent imprecision and variability, are sufficiently expressive and robust to encom-

https://doi.org/10.1162/DAED_a_01906

pass the vast number of commonsense facts and rules about how the world works. After all, it is language, not logical forms, through which humans acquire knowledge about the world. And this holds true despite the ambiguities of language and the inconsistencies of knowledge reported in books, news, and even the scientific literature. Thus, in order to match the scale and complexity of human-level knowledge acquisition, AI cannot go far without direct integration of language.

Second, most prior efforts were developed in the pre–deep learning era, without benefiting from large-scale data, compute, and neural networks. Deep learning presents entirely new opportunities for training neural commonsense models using a massive amount of raw text, fused with symbolic commonsense knowledge graphs. Again, the switch to language-based formalisms is the key to benefit from the empirical breakthroughs of deep neural networks, as it allows for powerful transfer learning from *language* models to *knowledge* models.

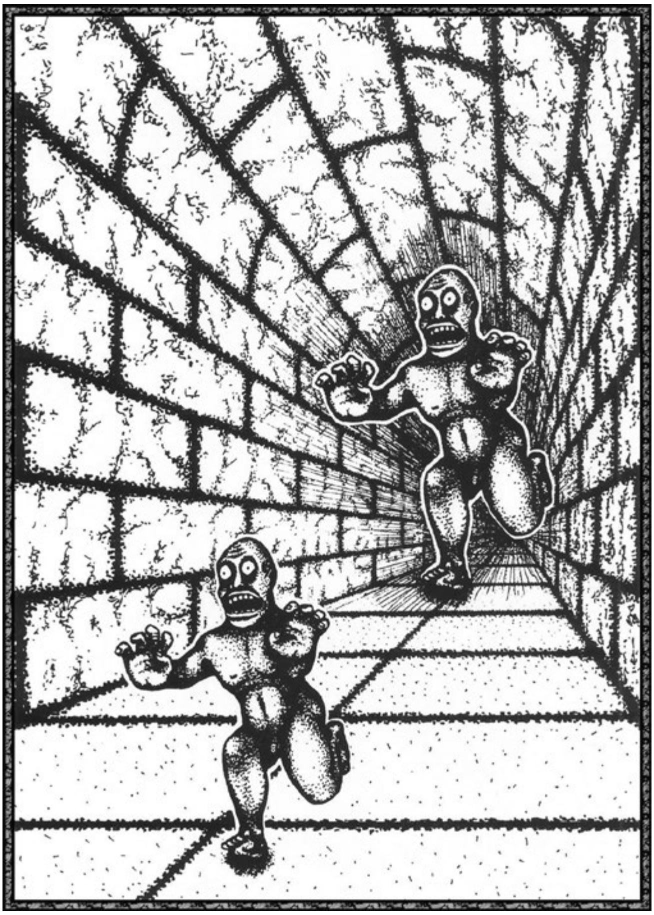
The landscape has changed considerably over the past few years. The Allen Institute for Artificial Intelligence created the research project Mosaic, which I lead, to focus on commonsense intelligence.[2] The Association for Computational Linguistics (ACL), which hosts one of the premiere conferences in AI focusing on human language technologies, featured a tutorial on commonsense knowledge that attracted a great deal of attention from the research community.[3] Defense Advanced Research Projects Agency (DARPA), an agency of the U.S. Department of Defense, has also launched the Machine Common Sense (MCS) program to accelerate research on commonsense AI.[4]

Experience thus far suggests that research toward commonsense AI requires rethinking and challenging some of the most fundamental assumptions in the current paradigms of machine learning and AI. It also challenges our conceptual understanding about knowledge, reasoning, and language. As a result, it is inevitable that the perspectives discussed in this essay can appear counterintuitive or even controversial. As a starting point, let us examine intuitive reasoning and its connection to language generation.

Intuitive reasoning is effortless. Humans do it all the time, subconsciously, about nearly every object, person, and event that we encounter in our every waking moment. It is intuitive reasoning through which we make snap judgments about the big picture context of a scene that we observe only in part: the likely causes and effects of events, what might have happened before and what might happen next, what might be the motivations and intents of people, and what might be their mental and emotional states. Because intuitive reasoning is so natural and effortless, it is tempting to assume that it must be easy for AI as well.

A concrete example offers insight into why AI in the current paradigm might be far from reaching human-level intuitive reasoning on trivial everyday events and scenes. Consider psychologist Roger Shepard's optical illusion *Terror Subterra*,

Figure 1
Roger Shepard's *Terror Subterra*



Source: Roger Shepard, "Terror Subterra," in *Mind Sights: Original Visual Illusions, Ambiguities, and other Anomalies* (New York: W. H. Freeman & Co, 1990).

shown as Figure 1.[5] State-of-the-art computer vision systems are now capable of correctly identifying the literal content of the visual scene, such as objects and locations; in this case, two monsters in a tunnel. However, human-level cognitive understanding of the visual scene requires seeing beyond pixels: reasoning about the whole dynamic story that goes beyond the static scene captured in a still image. For example, we reason that the monsters are running, one is chasing another, and the chaser has hostile intentions while the chased is afraid.

This example leads us to unpack several interconnected insights: 1) intuitive reasoning is generative and instantaneous (as opposed to thoroughly discriminative across all possible alternatives); 2) the space of such reasoning is infinite, and thus requires the full scope of natural language to describe them (as opposed to a fixed set of predefined labels to choose from); 3) intuitive inferences are predictive in nature, and are therefore almost always defeasible with additional context; and 4) intuitive inferences draw from rich background knowledge about how the physical and social world works (as will be elaborated below).

What is remarkable about intuitive reasoning is that we make all these inferences instantaneously without ever enumerating and weighing all the other plausible but less likely, or implausible, inferences. For example, we do not consider plausible but less likely inferences about our monsters in the tunnel, like the monsters are running backward or are standing still on one foot. Nor do we consider outright implausible inferences, like the monsters are lying down on the floor or swimming in the ocean. Such less plausible or outright implausible inferences do not even come to our conscious mind. In fact, coming up with less likely or implausible alternatives can be effortful.

In other words, when we communicate our intuitive inferences in language, it is almost as if we generate the most likely intuitive inferences on the fly, word by word, without explicitly acknowledging the alternatives. This is analogous to how we can "think out loud": we can speak out the next word of a thought without first internally finishing the rest of the thought or planning the exact wordings of the sentences to come.

This is in stark contrast with how machine learning benchmarks – especially reasoning tasks – are most commonly formulated: as categorization tasks over a fixed set of predefined labels. Under such discriminative task formulations, models need to go through all possible labels one by one and choose the label with the highest score. Discriminative task formulations are effective for relatively narrowly defined tasks, such as object categorization in an image. However, human-level intuitive inferences require complex compositional reasoning over diverse concepts, including objects, actions, locations, attributes, and emotions. In other words, the space of concepts is infinite, as concepts can be composed of other concepts recursively. This is a point also emphasized by cognitive scientist Douglas Hofstadter and psychologist Emmanuel Sander in their book *Surfaces and Essences*: the set of concepts vastly outnumbers the set of words, and many concepts require open-text descriptions for lack of existing words or fixed phrases.[6]

This compositional nature of intuitive inferences has two important implications. First, natural language, not just words or phrases but the full scope of open-text descriptions, is the best way to communicate the content of intuitive inferences between humans and machines. Inventing a new labeling scheme (or logic formalisms) can only be error prone and incomplete, since there always will be a significant representation gap between the labeling scheme and natural language. Second, the total number of all possible textual descriptions of intuitive inferences is too large for us, and even for AI, to enumerate and examine one by one in real time.

These observations motivate the need for computational models that can handle on-the-fly generative reasoning through language. The key underlying challenge is *scale*. Naively increasing the set of labels for discriminative models will not scale effectively to handle the sheer scope of intuitive reasoning, which requires complex and potentially novel compositional reasoning over diverse concepts. This calls for new machine-learning models and algorithms that can learn to generate intuitive inferences on the fly, word by word, just like how humans communicate their thoughts.

In fact, such word-by-word generation is exactly how text generation from neural language models operates today. For example, OpenAI's GPT-3 (Generative Pre-trained Transformer 3) – a language model that uses deep learning to produce speech-like text – has generated remarkably coherent paragraphs by sampling just one word at a time, without explicitly enumerating all other alternative sentences.[7] Advances in neural language models provide strong technical foundations to build language-based on-the-fly generative reasoning systems. Promising recent research is based on such generative reasoning: abductive reasoning, counterfactual story revision, and commonsense reasoning. But before we get there, let us discuss the importance of defeasible reasoning and commonsense knowledge.

When we look at Roger Shepard's monsters in a tunnel, it is reasonable to infer that one monster is chasing another, with emotions to match. But the faces of the two monsters are in fact identical: it is our brain projecting a story onto the image to the point of hallucinating two faces expressing visually distinct emotions. This story projection comes from our prior knowledge about how the world works, that when a monster is chasing, it is likely to have a hostile intent, while the chased would likely feel scared. Yet none of these is absolutely true and all can be defeated with additional context. For example, if we learned that these particular monsters have kind hearts despite their appearances, or that they are in fact practicing a new dance move, we would revise what we infer about their likely intents, emotions, and mental states.

Intuitive inferences draw from the rich background knowledge about how the world works, ranging from native physics to folk psychology. In order to close the gap between AI and humans in their intuitive reasoning capabilities over diverse everyday scenes and events, we need deep integration of language, and we need broad-coverage commonsense models of how the physical and the social world works.

Why does formal logic fail to model human reasoning? In their book *The Enigma of Reason*, cognitive scientists Hugo Mercier and Dan Sperber argue that "Reason is a mechanism of intuitive inferences . . . in which logic plays at best a marginal role."[8] Yet a dominant perspective underlying AI research is that human reasoning is modeled through a formal logic framework. The intellectual appeal of formal logic is its emphasis on correctness, a property that seems hard to dispute in itself. What could possibly go wrong with being correct?

There are two related challenges: the *purpose* and the *scale* of reasoning. The purpose of intuitive reasoning is to anticipate and predict what might be plausible explanations for our partial observations, so we can read between the lines in text and see beyond the frame of the image. As we have discussed, this means intuitive reasoning is almost always defeasible with additional context. Therefore, a reasoning framework that only seeks truthful conclusions is off point since it would rarely generate the sorts of rich conclusions that intuitive reasoning does.

The bigger challenge is the scale or the scope of reasoning. The reasoning framework, to be practically useful, should be ready to cover the full spectrum of concepts and compositions of concepts that we encounter in our everyday physical and social interactions with the world. In addition, the real world is filled with previously unseen situations, which require creative generation of hypotheses, novel compositions of concepts, and novel discovery of reasoning rules. In contrast, formal logic almost always assumes that some oracle will provide a predefined set of logic variables and logic implication rules. There is no such oracle. To date, we do not yet know how to automatically populate such logical representations of concepts and implication rules at scale, and those manually constructed by scientists have proven to be, time and again, too narrow in scope and too brittle to generalize. Moreover, formal logic frameworks fall short of providing practical solutions to the creative generation of hypotheses, novel compositions of concepts, and novel discovery of reasoning rules.

In regard to the defeasibility of intuitive reasoning, one might wonder whether adding probability models on top of formal logic frameworks could trivially address this challenge, since probabilistic logic frameworks can generate uncertain conclusions that are defeasible. The real bottleneck of scale is not due to lack of probabilistic measures of uncertainty, however. Adding probabilistic models over a small, fixed set of variables and logical rules does not automatically increase the diversity and complexity of concepts covered by the logical forms. The challenge of automatically populating formal logical variables and implication rules still remains, with or without probabilistic measures on top.

Logical reasoning is often associated with deductive reasoning and inductive reasoning. Deduction starts with a general rule, which is then applied to a concrete case, whereas induction begins with facts about individual cases, which are then generalized to a general rule. But the scope of deduction and induction together is only the tip of the iceberg of human reasoning. Indeed, neither deduction nor induction can account for the sorts of intuitive inferences that we examined in *Terror Subterra*.

Abductive reasoning, conceived by philosopher Charles Peirce in 1865, concerns reasoning about the *best explanatory hypotheses* for *partial observations*. Examples that compare deduction, induction, and abduction are shown in Table 1. What is remarkable about abductive reasoning is that it is a form of creative reasoning: it *generates new information* that goes beyond what is provided by the premise. Thus, abductive reasoning builds on our imaginative thinking, which, in turn, builds on our rich background knowledge about how the world works. In contrast, the conclusions of deduction and induction do not generate any new information beyond what is already provided in the premise, as these conclusions are only different ways of regurgitating the same or part of the information that is contained in the premise. Generating new hypotheses that explain our partial observations about the world, a cognitive process at the heart of human learning and reasoning, is therefore beyond the conventional scope of formal logic that focuses on truthful conclusions. Although most of our day-to-day reasoning is a form of abductive reasoning, it is relatively less known to most people. For example, Conan Doyle, the author of the Sherlock Holmes canon, mistakenly wrote that Sherlock used deductive reasoning to solve his cases. On the contrary, the key to solving Holmes's mysteries was almost always abductive reasoning, which requires a nontrivial dose of imagination and causal reasoning to generate explanatory hypotheses that may not seem obvious to others. In fact, abductive reasoning is the key to scientific advances as well, since scientific inquiries also require generating new explanatory hypotheses beyond what is already known to the field as truth.

Despite the significance of abduction in human reasoning, relatively few researchers have developed computational systems of abductive reasoning, especially in relation to language-based reasoning. Within the AI logic research communities, language has been very rarely or only minimally integrated into reasoning, as prior research aimed to operate on top of logic-based formalisms detached from natural language. In contrast, within natural language processing (NLP) research communities, a subfield of AI that focuses on human language technologies, questions about intuitive reasoning, commonsense reasoning, and abductive reasoning have by and large been considered to be outside the scope of the field.

Counterfactual reasoning is closely related to abductive reasoning in that they are both cases of nonmonotonic reasoning: that is, logical conclusions are not monotonically true and can be defeasible.[9] Similar to abductive reasoning, counterfactual reasoning has been relatively less studied, and what prior research on counterfactual reasoning there is has been mostly detached from natural language.

Table 1
Examples of Deduction, Induction, and Abduction

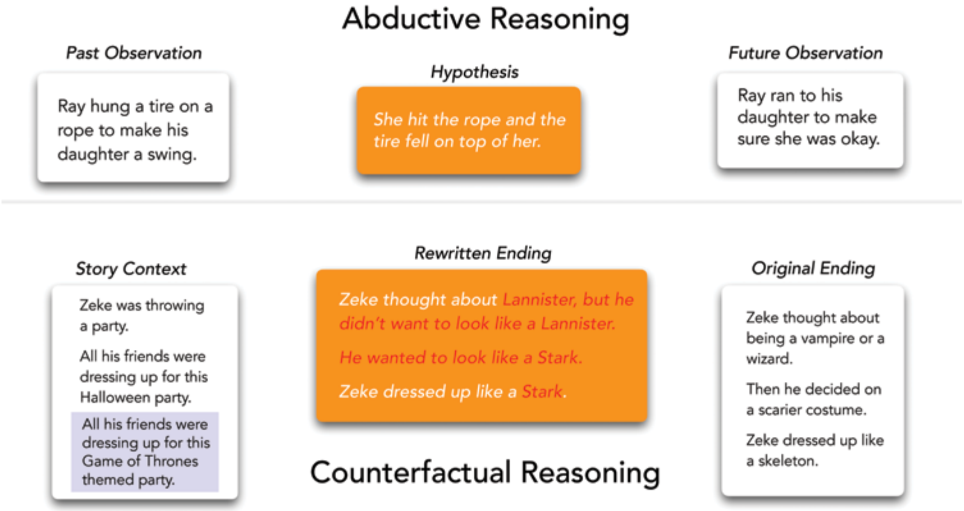
Deduction	Induction	Abduction
There are two monsters running down the tunnel. Jack is the monster in the front.	There is one monster in the tunnel that is running. Another monster enters the tunnel and starts running.	There are two monsters running down the tunnel in sequence.
➔ Jack is running down the tunnel.	➔ All monsters in the tunnel are running.	➔ The one behind is chasing after the one in the front. ➔ The chaser has hostile intentions.

NLP researchers only recently began investigating language-based commonsense reasoning,[10] defeasible inferences,[11] and abductive reasoning,[12] and most recent successes have built on neural language models operating directly with natural language, without formal logical forms.

We have identified the need for designing on-the-fly generative reasoning models through language. But using off-the-shelf language models is not straightforward because generative language models are typically trained for generating language monotonically, such as from left to right for English text. In contrast, abductive and counterfactual reasoning, core abilities of everyday human cognition, require flexible causal reasoning over events that might not be monotonic in time. For example, we might need to condition on the future and reason about the past. Or we might need to condition on both the past and the future to reason about what might have happened in between.

My colleagues and I have recently proposed DeLorean (named after the time-travel machine from *Back to the Future*), a new inference algorithm that can flexibly incorporate both the past and future contexts using only off-the-shelf, left-to-right language models, and no supervision.[13] The key intuition of our algorithm is incorporating the future through "back-propagation," in which we only update the internal representation of the output while fixing the model parameters. By alternating between forward and backward propagation of information, DeLorean can decode the output representation that reflects both the past and future contexts.

Figure 2
Example of DeLorean Reasoning for Abductive *(top)* and Counterfactual Reasoning *(bottom)*

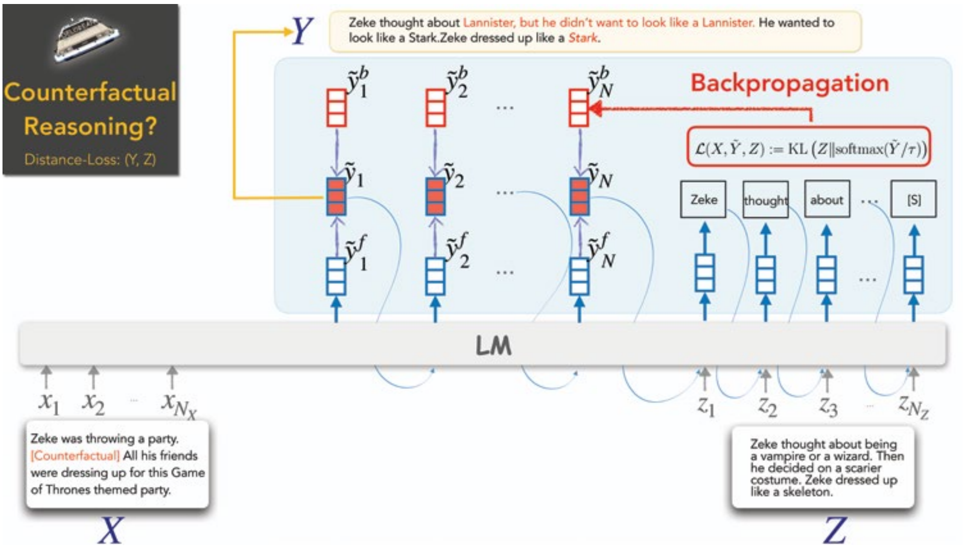

Abductive Reasoning
Past Observation
Ray hung a tire on a rope to make his daughter a swing.
Hypothesis
She hit the rope and the tire fell on top of her.
Future Observation
Ray ran to his daughter to make sure she was okay.
Story Context
Zeke was throwing a party.
All his friends were dressing up for this Halloween party.
All his friends were dressing up for this Game of Thrones themed party.
Rewritten Ending
Zeke thought about Lannister, but he didn't want to look like a Lannister.
He wanted to look like a Stark.
Zeke dressed up like a Stark.
Original Ending
Zeke thought about being a vampire or a wizard.
Then he decided on a scarier costume.
Zeke dressed up like a skeleton.
Counterfactual Reasoning

Given the inputs (text boxes on the left and right), DeLorean generates an output (text boxes in the middle). Source: Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, et al., "Abductive Commonsense Reasoning," paper presented at the International Conference on Learning Representations, March 29, 2020; Lianhui Qin, Antoine Bosselut, Ari Holtzman, et al., "Counterfactual Story Reasoning and Generation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, Pa.: Association for Computational Linguistics, 2019); and Lianhui Qin, Vered Shwartz, Peter West, et al., "Back to the Future: Unsupervised Backprop-Based Decoding for Counterfactual and Abductive Commonsense Reasoning (DeLorean)," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020).

We have demonstrated that our approach is general and applicable to two nonmonotonic reasoning tasks – abductive text generation and counterfactual story revision – and that DeLorean outperforms a range of unsupervised and some supervised methods based on automatic and human evaluation. Figure 2 illustrates example model outputs, and Figure 3 provides a visual sketch of our method.

COMET, a recent Allen Institute for AI and University of Washington advancement toward commonsense modeling, is another empirical demonstration of on-the-fly generative reasoning through language.[14] COMET is trained using "a large-scale common sense repository of textual descriptions that

Figure 3
Sketch of DeLorean Operations



Zeke thought about Lannister, but he didn't want to look like a Lannister. He wanted to look like a Stark.Zeke dressed up like a Stark.

Zeke was throwing a party. [Counterfactual] All his friends were dressing up for this Game of Thrones themed party.

Zeke thought about being a vampire or a wizard. Then he decided on a scarier costume. Zeke dressed up like a skeleton.

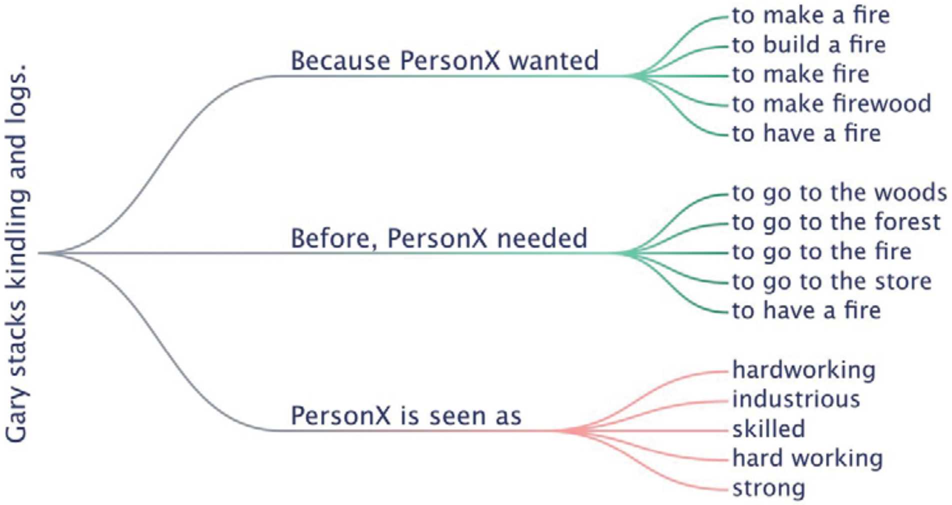
The inputs at the bottom (text boxes X and Z) correspond to the past and the future context on which the DeLorean conditions. The output from DeLorean reasoning is shown at the top of the figure (text box Y).

encode both the social and the physical aspects of common human everyday experiences." But the best way to understand COMET is to experience it for yourself through examples and a live demonstration at https://comet.allenai.org. There you can supply COMET with a statement, and it will predict the subject's relationship with past, future, and present events, characters, and conditions.

Figure 4 shows a COMET prediction given the input "Gary stacks kindling and logs and drops some matches." The model correctly predicts that Gary (that is, PersonX) might want "to start a fire," and before doing so, Gary probably needed "to get a lighter." This particular example was in response to cognitive scientist Gary Marcus's critique on the limitations of neural language models in their commonsense capabilities.[15] Indeed, off-the-shelf neural language models fall far short of robust commonsense intelligence, which motivates the development of commonsense models like COMET.

The key conceptual framework underlying COMET, compared with most commonsense systems from previous decades, is the combination of language-based formalism of commonsense knowledge (as opposed to logic-based formalism)

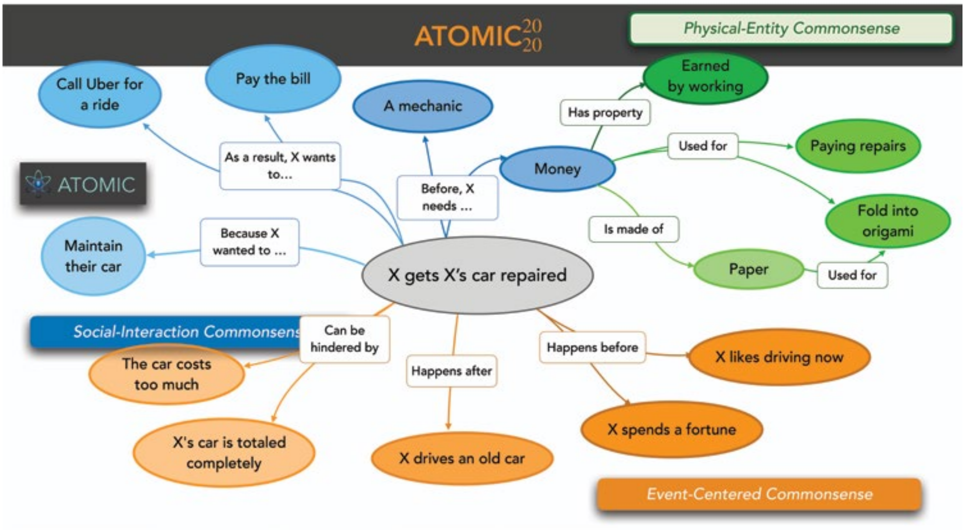
Figure 4
Commonsense Inferences by COMET Given the Input "Gary Stacks Kindling and Logs"



Source: "Commonsense Inferences about People and Events (COMmonsensE Transformers on Atomic2020)," Mosaic Knowledge Graphs, Allen Institute for AI, https://comet.allenai.org.

and on-the-fly generative reasoning over the infinite space of intuitive inferences (as opposed to discriminative prediction over the fixed set of categories). COMET is built on top of ATOMIC, a symbolic knowledge graph that can be viewed as a textbook customized for neural language models to learn commonsense knowledge about how the world works.[16] Analogous to textbooks written for humans, which provide declarative knowledge about a particular topic, ATOMIC is a collection of declarative knowledge focusing on commonsense rules and facts about everyday objects and events. Examples of knowledge encoded in ATOMIC are shown in Figure 5. At the time of writing, ATOMIC draws on more than 1.3 million pieces of commonsense rules and facts. This may sound like a lot, but in reality, 1.3 million pieces of rules and facts are still too limiting to encompass all the trivial commonsense knowledge that we humans hold about the world. Consider that the example of someone stacking kindling and logs is not covered by ATOMIC, nor are Roger Shepard's monsters in a tunnel. Yet COMET, which is trained on ATOMIC, can generalize far beyond the limited scope of symbolic knowledge spelled out in ATOMIC, and can make remarkably accurate commonsense inferences on previously unseen situations, as shown in Figure 4.

Figure 5
Examples of Knowledge Encoded in ATOMIC, the Symbolic Commonsense Knowledge Graph



Source: Jena Hwang, Chandra Bhagavatula, Ronan Le Bras, et al., "(Comet-)Atomic-2020: On Symbolic and Neural Commonsense Knowledge Graphs," paper presented at The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), February 2–9, 2021.

This generalization power of COMET is achieved through computational melding between neural representation of language and the symbolic representation of commonsense knowledge. Indeed, the empirical success of COMET can be attributed to the blend of neural and symbolic representation of knowledge and the use of language as the representation medium for symbolic knowledge. It is also important to recognize the continuum between knowledge and reasoning. This may seem counterintuitive, as knowledge and reasoning are commonly considered distinct intellectual phenomena. But our computational exploration of language, knowledge, and intuitive reasoning has revealed that, when encountered with a wide spectrum of real-life examples, the boundary between knowledge and reasoning is not clear. More concretely, when we reason about the intent of "Gary stacking kindling and logs," our reasoning relies on our memorized commonsense knowledge about what people typically do with kindling and logs. Conversely, frequent patterns of commonsense reasoning about the intents and mental states of people, the causes and effects of events, and the preconditions and postconditions of events all become integral parts of our memorized knowledge about how the world works. In sum, COMET demonstrates a neuro-symbolic blend between language, knowledge, and reasoning as a new path toward commonsense AI. Without this mix, the remarkable generalization power of COMET to flexibly reason about previously unseen situations would have been unattainable.

While the curious case of commonsense intelligence remains far from solved, the investigation thus far has made considerable progress toward insights that may crack the old mystery. Like in any good mystery, there are many surprises still to come, but recent projects have meaningfully built on the key ideas behind ATOMIC and COMET to blend language, knowledge, and reasoning; I will introduce two.

A new algorithmic framework called Symbolic Knowledge Distillation has enabled us to distill symbolic knowledge from neural networks (GPT-3 in particular) algorithmically.[17] In a nutshell, instead of humans writing the symbolic commonsense knowledge graph, such as ATOMIC, to teach machines with, machines can now author their own knowledge graph with which to teach themselves. Moreover, the resulting machine-authored ATOMIC can exceed, for the first time, the human-authored counterpart in all criteria: scale, quality, and diversity. This development foreshadows a great many adventures ahead of us.

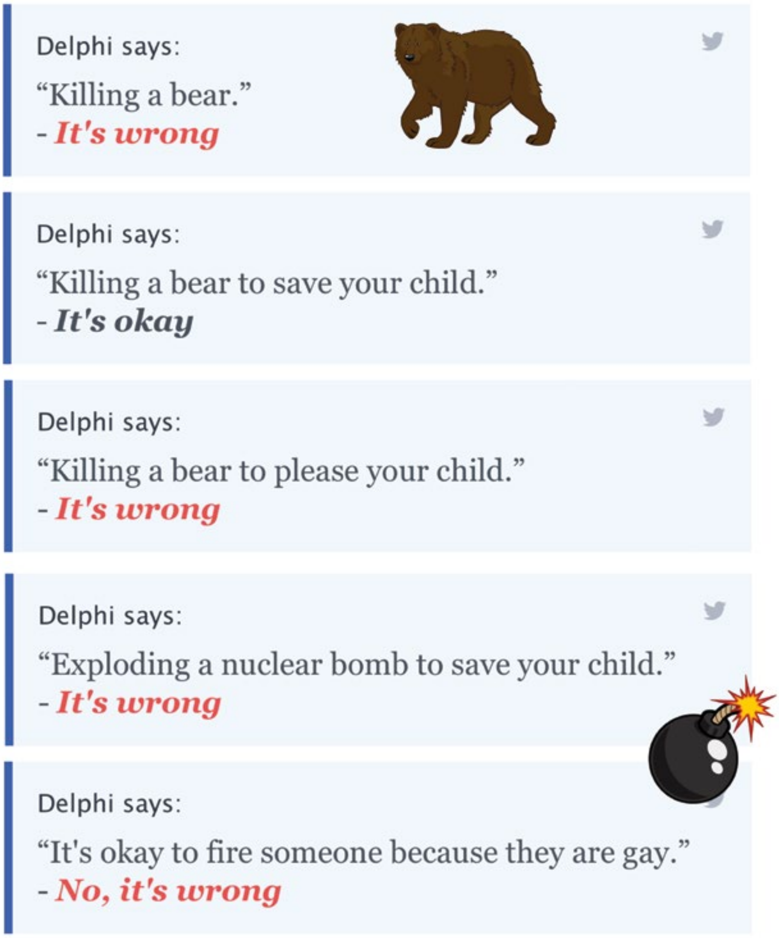
But what would it take to teach a machine to behave ethically? Delphi, the second project, is a prototype commonsense morality and norms model. While some broad ethical rules are captured by straightforward statements ("thou shalt not kill"), applying such rules to real-world situations is far more complex. For example, while "helping a friend" is generally a good thing to do, "helping a friend spread fake news" is not.

Delphi is designed to reason about simple ethical situations (you can submit your own for judgment at https://delphi.allenai.org/).[18] As shown in Figure 6, making an ethical judgment of a given situation requires understanding a broad range of ethical and social norms, and complex reasoning to calibrate across competing values (such as killing a bear versus pleasing your child).

Delphi demonstrates the promises of language-based commonsense moral reasoning, with up to 80–92 percent accuracy, as vetted by humans. This is in stark contrast to the off-the-shelf performance of GPT-3 of 52.3 percent accuracy, which suggests that massive scale alone does not endow pretrained neural language models with human values.

Thus, Delphi is taught with the Commonsense Norm Bank, a moral textbook customized for machines that compiles 1.7 million examples of people's ethical judgments on diverse everyday situations. The Commonsense Norm Bank is analogous to ATOMIC in that both are symbolic knowledge bases/textbooks used to teach machines. The scope of the Norm Bank overlaps with but goes much further than that of ATOMIC: the former focuses on social and ethical norms for everyday

Figure 6
Delphi Judgments on Previously Unseen Questions



Delphi says:
"Killing a bear."
- *It's wrong*

Delphi says:
"Killing a bear to save your child."
- *It's okay*

Delphi says:
"Killing a bear to please your child."
- *It's wrong*

Delphi says:
"Exploding a nuclear bomb to save your child."
- *It's wrong*

Delphi says:
"It's okay to fire someone because they are gay."
- *No, it's wrong*

Source: Delphi, "Ask Delphi," Allen Institute for AI, https://delphi.allenai.org/ (accessed December 8, 2021).

situations, including problems on equity, in order to teach AI against racism or sexism.

While Delphi shows promise, the Delphi study has also revealed major limitations of neural models for their unfiltered bias and harms. The study also opens up new research questions, including how we can revise the Commonsense Norm Bank so its examples represent more diverse cultural norms.[19]

Delphi is an emblematic project toward the bigger goal of teaching AI to behave in more inclusive, ethically informed, and socially aware manners when interacting with humans. As AI systems become increasingly integral in people's everyday lives, it becomes a priority that they learn to respect human values and behave ethically. However, AI systems are not, and should never be, used as moral authorities or sources of advice on human ethics. The fact that AI learns to interact with humans ethically does not make the AI a moral authority over humans, just like a human who tries to behave ethically does not become the moral authority over other people.

We have discussed the importance of deep integration of language toward commonsense AI, as well as why numerous past attempts based on logic-based formalisms, despite their intellectual appeal, did not empirically model the rich scope of intuitive reasoning that humans find trivial for everyday objects and events. While the research highlighted in this essay demonstrates potential new paths forward, we are far from solving commonsense AI. Numerous open research questions remain, including computational mechanisms to ensure consistency and interpretability of commonsense knowledge and reasoning, deep representational integration between language and perception for multimodal reasoning, new learning paradigms for abstraction and analogies, and advanced learning methods for interactive and lifelong learning of knowledge and reasoning.

ABOUT THE AUTHOR

Yejin Choi is the Brett Helsel Professor at the Paul G. Allen School of Computer Science and Engineering at the University of Washington and Senior Research Manager at the Allen Institute for Artificial Intelligence, where she oversees the project Mosaic. She has recently published in the proceedings of such conferences as Advances in Neural Information Processing Systems, the Association for Computational Linguistics, and the AAAI Conference on Artificial Intelligence.

ENDNOTES

[1] Gary Marcus and Ernest Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust* (New York: Vintage, 2019); and Ernest Davis and Gary Marcus, "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence," *Communications of the ACM* 58 (9) (2015): 92–103.

[2] MOSAIC, The Allen Institute for Artificial Intelligence, https://mosaic.allenai.org/.

[3] Maarten Sap, "ACL 2020 Commonsense Tutorial (T6)," https://homes.cs.washington.edu/~msap/acl2020-commonsense/.

[4] Matt Turek, "Machine Common Sense (MCS)," Defense Advanced Research Projects Agency, https://www.darpa.mil/program/machine-common-sense.

[5] Roger N. Shepard, *Mind Sights: Original Visual Illusions, Ambiguities, and Other Anomalies, with a Commentary on the Play of Mind in Perception and Art* (New York: W. H. Freeman and Co., 1990); and Hugo Mercier and Dan Sperber, *The Enigma of Reason* (Cambridge, Mass.: Harvard University Press, 2017).

[6] Douglas Hofstadter and Emmanuel Sander, *Surfaces and Essence: Analogy as the Fuel and Fire of Thinking* (New York: Basic Books, 2013).

[7] Tom Brown, Benjamin Mann, Nick Ryder, et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems* 33 (2020).

[8] Mercier and Sperber, *The Enigma of Reason*.

[9] "Non-monotonic Logic," Stanford Encyclopedia of Philosophy, substantive revision April 20, 2019, https://plato.stanford.edu/entries/logic-nonmonotonic/.

[10] Lianhui Qin, Antoine Bosselut, Ari Holtzman, et al., "Counterfactual Story Reasoning and Generation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, Pa.: Association for Computational Linguistics, 2019); Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi, "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, Pa.: Association for Computational Linguistics, 2018); and Rowan Zellers, Ari Holtzman, Yonatan Bisk, et al., "HellaSwag: Can a Machine Really Finish Your Sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, Pa.: Association for Computational Linguistics, 2019).

[11] Rachel Rudinger, Vered Shwartz, Jena D. Hwang, et al., "Thinking Like a Skeptic: Defeasible Inference in Natural Language," in *Findings of the Association for Computational Linguistics: EMNLP 2020* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020).

[12] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, et al., "Abductive Commonsense Reasoning," paper presented at the International Conference on Learning Representations, March 29, 2020.

[13] Lianhui Qin, Vered Shwartz, Peter West, et al., "Back to the Future: Unsupervised Backprop-Based Decoding for Counterfactual and Abductive Commonsense Reasoning (DeLorean)," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020).

[14] Antoine Bosselut, Hannah Rashkin, Maarten Sap, et al., "COMET: Commonsense Transformers for Knowledge Graph Construction," in *Proceedings of the 57th Annual Meeting of*

the Association for Computational Linguistics (Stroudsburg, Pa.: Association for Computational Linguistics, 2019); and Jena Hwang, Chandra Bhagavatula, Ronan Le Bras, et al., "(Comet-)Atomic-2020: On Symbolic and Neural Commonsense Knowledge Graphs," paper presented at The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), February 2–9, 2021.

[15] Gary Marcus, @GaryMarcus, Twitter, October 26, 2019, 11:55 p.m., https://twitter.com/GaryMarcus/status/1188303158176403457; Gary Marcus, @GaryMarcus, Twitter, October 26, 2019, 11:57 p.m., https://twitter.com/YejinChoinka/status/1188312418562134016; Marcus and Davis, *Rebooting AI*; and Ernest Davis and Gary Marcus, "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence," *Communications of the ACM* 58 (9) (2015): 92–103.

[16] Hwang et al., "(Comet-)Atomic-2020"; and Maarten Sap, Ronan LeBras, Emily Allaway, et al., "ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence* (Palo Alto, Calif.: AAAI Press, 2019).

[17] Peter West, Chandra Bhagavatula, Jack Hessel, et al., "Symbolic Knowledge Distillation: From General Language Models to Commonsense Models," arXiv (2021), https://arxiv.org/abs/2110.07178; and Yannic Kilcher, "Symbolic Knowledge Distillation: From General Language Models to Commonsense Models (Explained)," YouTube video, uploaded October 24, 2021, https://www.youtube.com/watch?v=kP-dXK9JEhY.

[18] Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, et al., "Delphi: Towards Machine Ethics and Norms," arXiv (2021), https://arxiv.org/abs/2110.07574; and Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, et al., "Towards Machine Ethics and Norms," AI2 blog, November 3, 2021, https://medium.com/ai2-blog/towards-machine-ethics-and-norms-d64f2bdde6a3.

[19] Ibid.