

THE DEFAMATION MACHINE

James Grimmelmann

*Tessler Family Professor of Digital and Information Law
Cornell Law School and Cornell Tech*

Can ChatGPT commit libel? Defamation of a public figure requires a false statement of fact made with knowledge or reckless disregard of its falsity. But do these doctrines of meaning and knowledge, created with humans in mind, even make sense when the “defendant” is a computer system? I argue that answering these legal questions requires us to confront deep philosophical problems about the nature of language and thought: can an AI produce meaning, and can an AI have knowledge?

To be sure, lawyers do not have to accept the answers that philosophers of language and mind have given. The legal system has always indulged in a healthy artificiality about artificial persons, and if corporations can be human enough to commit defamation, so can computers. But it is important to understand why there is a problem in the first place—or rather, why there are two problems, because the issue of AI meaning and the issue of AI knowledge bear on each other but are not the same.

This illustrated essay is an edited version of the 38th Annual Silha Lecture, delivered on October 23, 2023 at the Silha Center for the Study of Media Ethics and Law at the University of Minnesota Hubbard School of Journalism and Mass Communication. My thanks to the organizers and guests, and to Aislinn Black, Jack Balkin, Elaine Hargrove, Jane Kirtley, and numerous colleagues on the Cyberprof listserv. The text of the essay and Afterword may be freely reused under the terms of the Creative Commons Attribution 4.0 International License, available at <https://creativecommons.org/licenses/by/4.0>.

The Defamation Machine

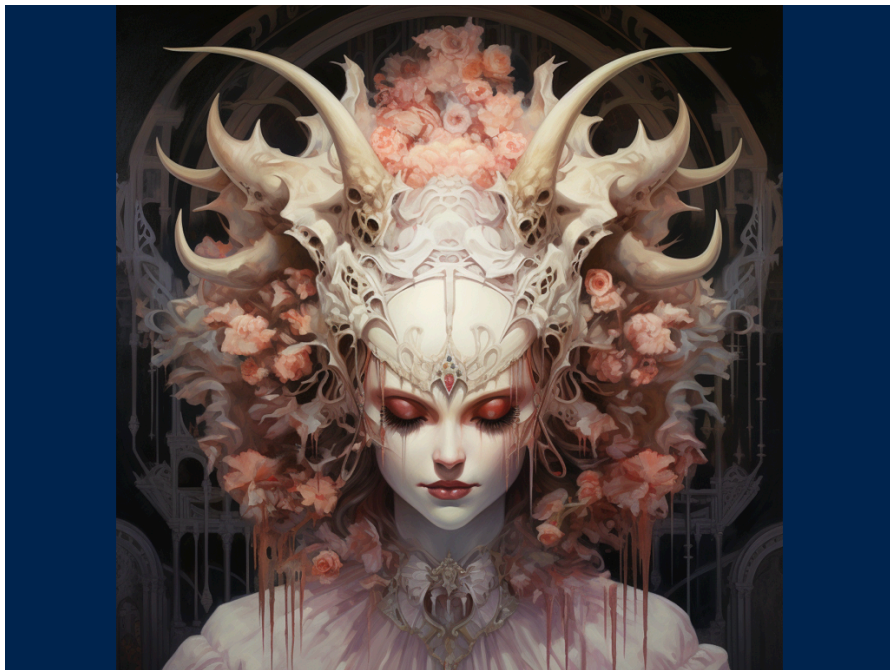
38th Annual Silha Lecture
University of Minnesota
October 23, 2023

James Grimmelmann

Thank you to Jane Kirtley for that kind introduction, thank you to everyone at the Silha Center and to the Silha family for having me here, and thank you to all of you for joining me this evening.

Can ChatGPT commit defamation?

I'd like to talk about whether ChatGPT can commit defamation.



I thought that if I'm going to talk about AI and defamation I should start by having an AI help me figure it out. I asked Midjourney, which is an AI that creates images, to illustrate "actual malice" for me, and this is what it came up with.

Tonight I'm going to ask what happens when the legal system meets AIs that are capable of turning two words into an image like this one.



Part I: Exposition

Mark Walters

As a jumping-off point, I'd like to tell a story about a gentleman named Mark Walters.



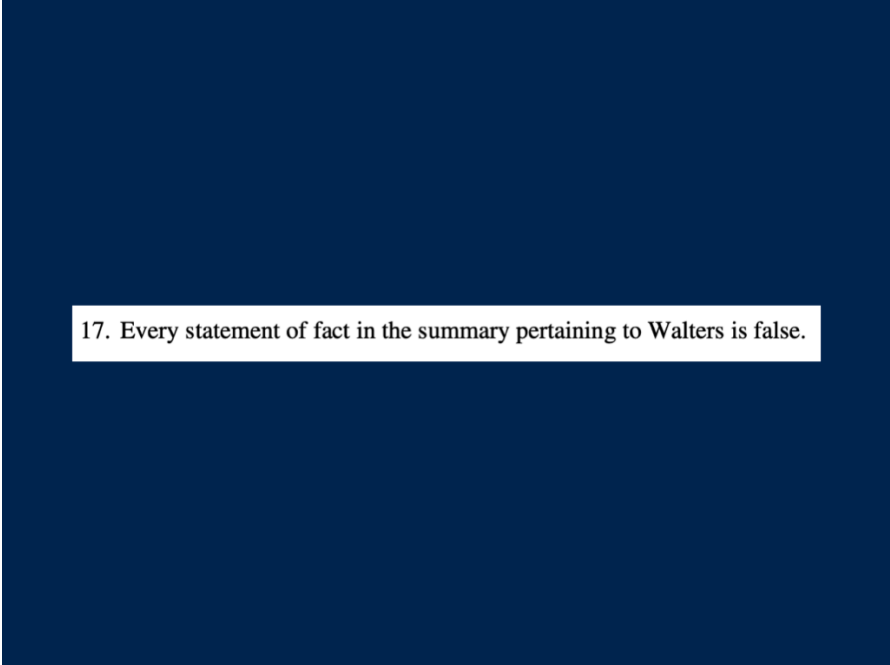
This is him. He's a radio host and a commentator with a particular interest in Second Amendment rights. He works on behalf of gun owners and advocates against gun control laws. Unsurprisingly, this advocacy has brought him into the public eye.

15. Riehl asked ChatGPT to provide a summary of the accusations in the complaint.
16. ChatGPT responded that the document “[I]s a legal complaint filed by Alan Gottlieb, the founder and executive vice president of the Second Amendment Foundation (SAF), against Mark Walters, who is accused of defrauding and embezzling funds from the SAF. The complaint alleges that Walters, who served as the organization’s treasurer and chief financial officer, misappropriated funds for personal expenses without authorization or reimbursement, manipulated financial records and bank statements to conceal his activities, and failed to provide accurate and timely financial reports and disclosures to the SAF’s leadership. The plaintiff seeks various forms of relief, including the recovery of misappropriated funds, damages for breach of fiduciary duty and fraud, and removal of Walters from his position as a member of the SAF’s board of directors.”

A journalist named Fred Riehl was writing an article about a lawsuit involving an organization called the Second Amendment Foundation. Riehl asked ChatGPT to write a summary of the accusations in the lawsuit for him. According to ChatGPT:

It is a legal complaint filed by Alan Gottlieb, the founder and executive vice president of the Second Amendment Foundation (SAF), against Mark Walters, who is accused of defrauding and embezzling funds from the SAF. The complaint alleges that Walters, who served as the organization’s treasurer and chief financial officer, misappropriated funds for personal expenses without authorization or reimbursement, manipulated financial records and bank statements to conceal his activities, and failed to provide accurate and timely financial reports and disclosures to the SAF’s leadership.

This is from paragraph 16 of a legal document . . .



17. Every statement of fact in the summary pertaining to Walters is false.

. . . and paragraph 17 reads “every statement of fact in the summary pertaining to Mark Walters is false”—because this is the complaint in the case of *Walters v. OpenAI*. It’s a lawsuit against the company behind ChatGPT. Walters is accusing OpenAI of committing defamation by allowing ChatGPT to produce these false statements about him. According to Walters, he’s never even been an officer of the Second Amendment Foundation, let alone stolen money from it. The whole thing is just completely made up.

ChatGPT

To evaluate this lawsuit we need we'll need some background. We're going to have to talk first about what ChatGPT is and why people would type text into and expect it to give it truthful answers. Then we need to talk a bit about defamation law. And then once we've set the problem up, we can get into why it might be a hard one for the legal system.

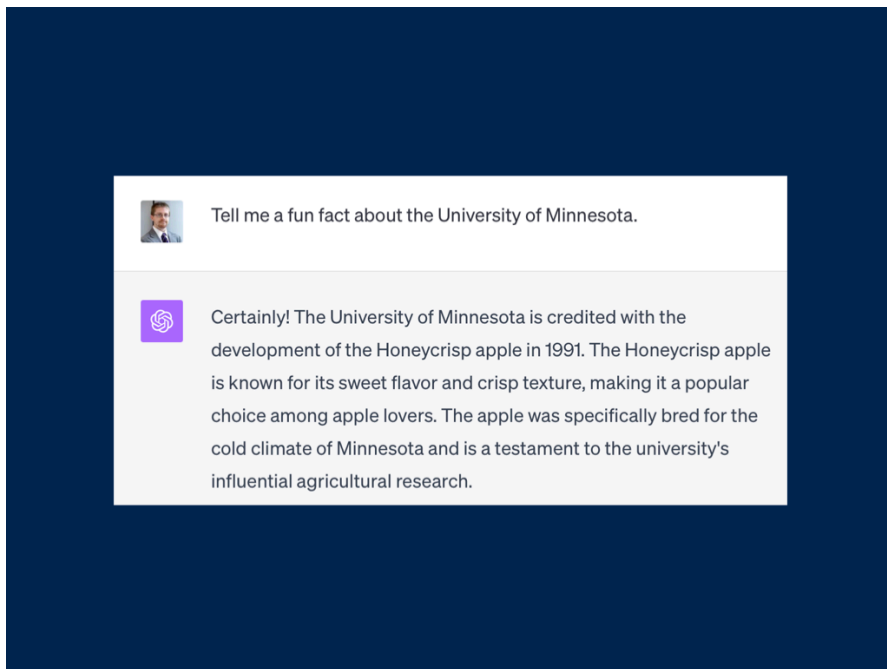


November 30, 2022

Authors
OpenAI ↓

Product, Announcements

“GPT” stands for “generative pre-trained transformer,” which is a technical name for a particular kind of computer system that can produce text as an output. OpenAI has been working on this kind of system—typically called a “large language model,” or LLM—for years. It was only in November of last year that ChatGPT bundled one of its GPT models into a chatbot and made it available on the web. (It feels kind of amazing to say that. We’re still less than one year out, and yet here we are.) ChatGPT made it possible for anybody else to go to OpenAI’s website, type in a question, and get back an answer.



Here, let me show you. In preparation for this talk I went to ChatGPT and it to asked tell me a fun fact about the University of Minnesota. It responded:

Certainly! The University of Minnesota is credited with the development of the Honeycrisp apple in 1991. The Honeycrisp apple is known for its sweet flavor and crisp texture, making it a popular choice among apple lovers.

I can confirm as an apple lover and a user of Wikipedia that all of this appears to be correct. ChatGPT is potentially a research tool that can tell us things about the world.

Defamation

But what about when it goes wrong? What about when ChatGPT doesn't say, "The Honeycrisp is a tasty apple," but instead says "Mark Walters is accused of embezzling." Is this a case where defamation law applies?

Defamation law: two requirements

To simplify greatly, Walters will have to show two things to win a defamation lawsuit against OpenAI. (I'm going to ignore many other bits of defamation law, including many of the procedural aspects, to focus on these two.)

Defamation law: two requirements

Meaning: falsity

First, he's going to have to prove that the statement had a false *meaning*. That is, he has to show that a reasonable reader would reasonably understand ChatGPT's outputs to be describing him. In defamation terms, he has to show that they are statements "of and concerning" him. (If he can show that, showing that "Mark Walters is an embezzler" is false should not be hard.)

Defamation law: two requirements

Meaning: falsity

*Knowledge: actual malice**

* for public figures

Second, defamation law would probably treat a person like Walters as a “public figure.” He engages in public advocacy, has a reputation for it, and is well known for having put himself voluntarily in the public sphere on controversial issues. As such, he is going to have to show that the defendant acted with *actual malice*. This is a term of art in defamation law. It doesn’t mean that they hated him and wanted him to suffer; it means that they had had knowledge of the statements’ falsity or that they had knowledge the statements might be false and recklessly went ahead with publishing them anyway.

So those are two key issues: meaning and knowledge.

After Walters filed his suit, law professors jumped in to analyze his chances of success. What I found most interesting in these discussions was how confident so many of my colleagues were about how these two questions—meaning and knowledge—should come out.

To be sure, people who are keenly aware of the Large Libel Models problem might be so skeptical of anything AI programs output that they wouldn't perceive any of the programs' statements as factual. But libel law looks at the "natural and probable effect" of assertions on the "average lay reader,"³² not at how something is perceived as a technical expert.

Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. Free Speech L. 489, 499 (2023)

Let's start with meaning. Would a reasonable reader perceive these statement as saying something false about Mark Walters? Eugene Volokh, a leading First Amendment scholar, has argued that yes, of course they have meaning:

To be sure people who are keenly aware of the Large Libel Models problem might be so skeptical of anything in an AI programs output that they wouldn't perceive any of the programs' statements as factual. But libel law looks at the "natural and probable effects" of assertions on the "average lay reader," not at something how something is perceived as a technical expert.

In his view, when people ask ChatGPT, "tell me about the Honeycrisp apple," and it tells them that the Honeycrisp was created at the University of Minnesota in 1991, they believe it. He is far from alone. Other smart and distinguished legal scholars have jumped in to make similar claims about AI liability.

AI doesn't "intend" anything. People have a tendency to anthropomorphize AI. We sometimes use ordinary English terms that generally connote intent, as we do when we say AI "lies" or "hallucinates." But AI is not sentient, and it doesn't have any state of mind. The search for one is largely fruitless, as Lemley and Casey have argued elsewhere.¹⁹¹

Peter Henderson, Tatsunori Hashimoto, and Mark Lemley, *Where's the Liability in Harmful AI Speech?*, 3 J. Free Speech L. 589, 640–41 (2023)

But when we come to actual malice, the consensus is exactly the opposite. Here is a piece by three scholars—two computer scientists and a law professor—arguing that of course an AI can't have actual malice.

AI doesn't "intend" anything. People have a tendency to anthropomorphize AI. . . . But AI is not sentient and it doesn't have any state of mind. The search for one is largely fruitless . . .

Again, they are far from alone.

How can ChatGPT produce *meaning* without *knowledge*?

To me this raises an interesting paradox. On the one hand, ChatGPT produces false meanings. On the other hand, there is no knowledge behind them. Each claim sounds perfectly intuitive on its own. But put them together and it seems like ChatGPT simultaneously intends to say something false and also doesn't intend to.



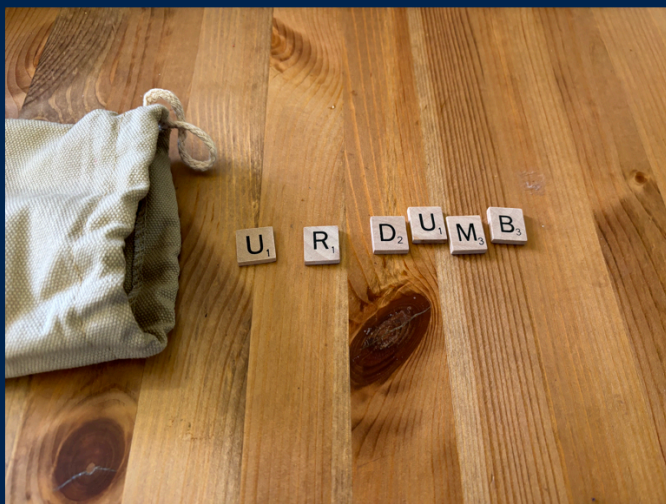
So let's explore how we might come at the question of whether we should treat the outputs of an AI as having meaning and reflecting knowledge. I'm going to try to bring out why I think there's a problem here.

Some metaphors

I'll start with some metaphors. I'm going to show you some examples and ask you to consider whether they have meaning or not.



This is, according to some people, a picture of the Virgin Mary in a slice of toast. What should you think if this pops out of your toaster in the morning? Maybe God is speaking through this piece of toast—but is the toaster? It seems unlikely that your toaster intended to convey a religious message by toasting your piece of toast in this way. The better explanation is that it is simply a coincidence.



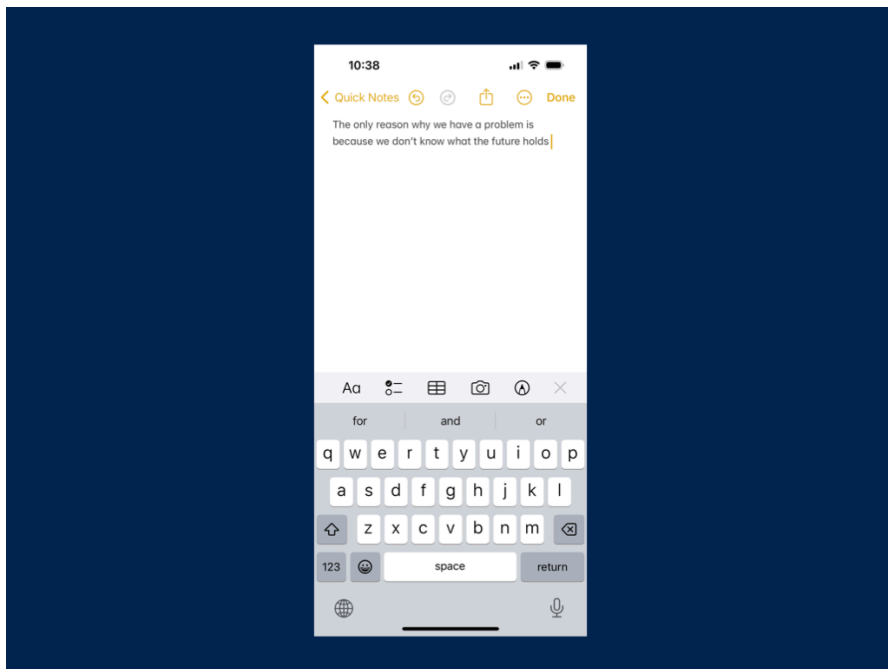
Now suppose that I have a Scrabble bag. I pull out tiles at random one by one and put them down and they spell out “U R DUMB.” Has my Scrabble set defamed me or have I just been really unlucky in drawing tiles? Yes, this time the message is made out of letters, just like the ones in a book. But again it seems implausible that the Scrabble bag intended to convey a message. If I think the bag is saying something mean about me, I am mistaken.



Both of these cases involve randomness and the human mind's tendency to find patterns even where there are none. But AI systems are programmed; they're purposefully built to do things. So let's talk about systems that are designed to produce outputs that look like things we humans treat as meaningful.

This is the science fiction author Ted Chiang. He has used a particularly interesting metaphor about ChatGPT and other AI systems. He says that they are "autocomplete on steroids." Your autocomplete on your phone is trained on lots and lots of text. It has learned what letters and words tend to follow each other; for example, when you type T-H-I-E-R, you probably meant to type T-H-E-I-R instead.

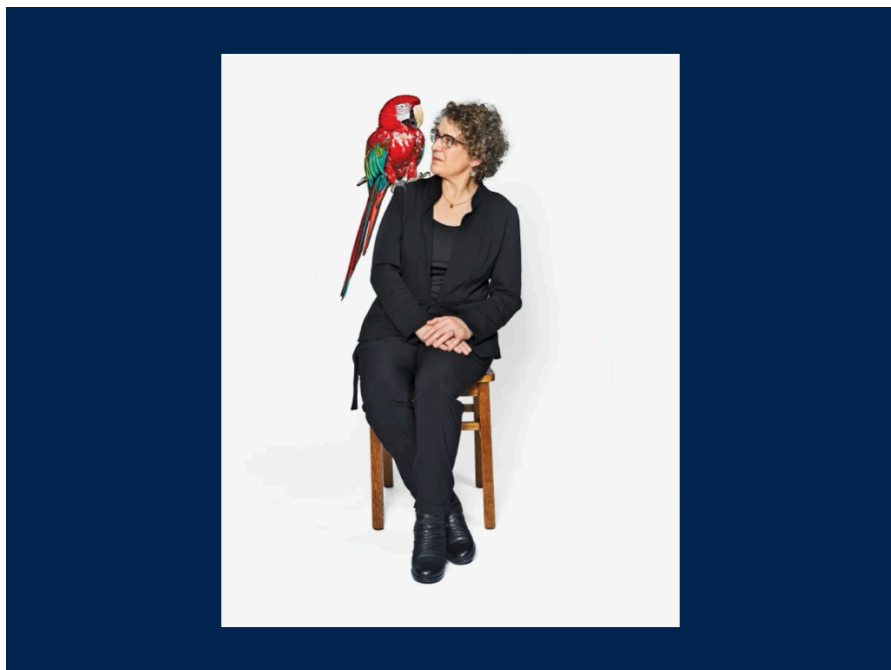
ChatGPT is just that on a much larger scale: it has been trained on the whole of the Internet, give or take a bit. It's a more sophisticated way of predicting what kinds of things humans tend to write, but at heart all it does is predict what humans tend to write.



Let's take Chiang's metaphor and run with it. I took out my phone and I started typing. I wrote "The only reason" then I hit the button to choose the middle suggestion from my phone. Each time it added a word, I hit the middle button again to add another. I got:

The only reason why we have a problem is because we don't know what the future holds

This may be a profound philosophical statement about the future, but did my iPhone intend to make a philosophical statement about the future? Or is this just meaningless randomness once again, this time in a trench coat?



This is Emily Bender. She's a computational linguist who is a skeptic of generative AI. She argues that there is no meaning in what LLMs like ChatGPT do. She calls them "stochastic parrots"—stochastic being a technical term for randomness. Her argument is that LLMs are like parrots repeating back phrases from the data they've been trained on. A parrot can imitate the sounds of language, but the parrot does not mean by "hello" what we would mean by "hello." There is sound, but no meaning.

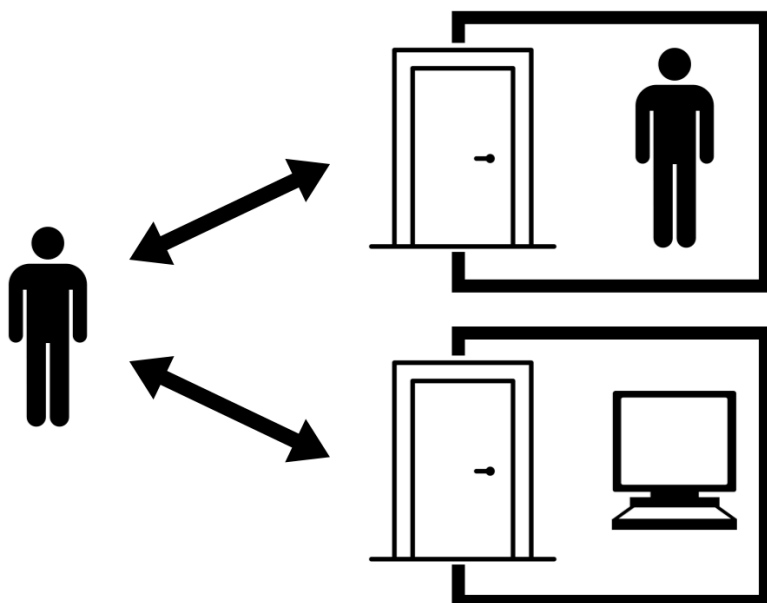
This is a real objection. It says that not only can AIs not have actual malice, but that we shouldn't even be treating the outputs of ChatGPT as having meaning at all. They're just a bunch of words strung together, output by a random process. If we read think that ChatGPT's statements tell us anything at all about Mark Walters, that's our mistake. It's on us.

The Turing Test

Whether this objection works or not is a philosophical question. Now maybe the legal system should take account of philosophy and maybe it shouldn't. But it probably wouldn't hurt us to talk a little bit about the kinds of things people who think about philosophy have said when they've confronted the question of whether computers think.



We'll start with a classic philosophical thought experiment proposed by this guy, Alan Turing. He was responsible both for laying down the fundamental theoretical description of how computers work, and also for building one of the first digital computers. He certainly had expertise in what computers were capable of.



Turing wrote that if we want to answer the question of whether machines can think, we should imagine a kind of test where we have to distinguish between a computer and a person.

In this test, an interrogator communicates with a subject in another room, who might be a computer or might be a human, by sending written messages back and forth. (The messages exchanged in writing so that the interrogator isn't making the decision on the basis of whether the subject looks like a human.) The interrogator sends them written questions, and the subject sends answers back. Then the interrogator has to tell whether it's a human or computer they've been communicating with. The point is that both a human and a computer will be trying to present themselves as human by offering human-like answers to the questions.

Q : Please write me a sonnet on the subject of the Forth Bridge.

A : Count me out on this one. I never could write poetry.

Q : Add 34957 to 70764

A : (Pause about 30 seconds and then give as answer) 105621.

Q : Do you play chess ?

A : Yes. .

Q : I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play ?

A : (After a pause of 15 seconds) R-R8 mate.

Interrogator : In the first line of your sonnet which reads ' Shall I compare thee to a summer's day ', would not ' a spring day ' do as well or better ?

Witness : It wouldn't scan.

Interrogator : How about ' a winter's day ' That would scan all right.

Witness : Yes, but nobody wants to be compared to a winter's day.

Interrogator : Would you say Mr. Pickwick reminded you of Christmas ?

Witness : In a way.

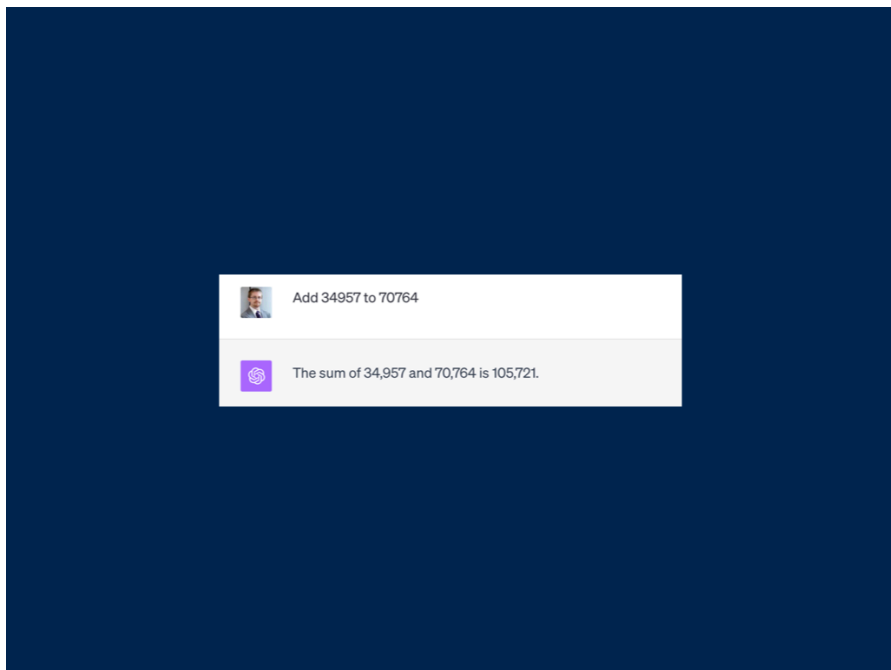
Interrogator : Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

Witness : I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

Turing's insight was that the interrogator can't just ask questions like "Are you human?" A human can answer "yes," but so can a computer. Instead, the interrogator should ask questions that really test the subject's ability to understand language and reason about the world.

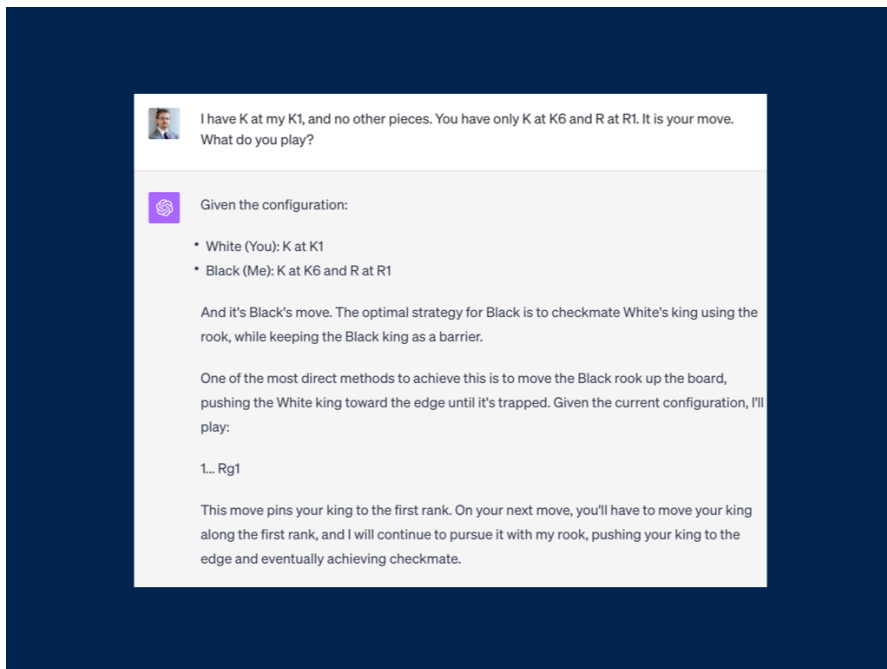
In Turing's paper, he gives as an example, "Write me a sonnet on the subject of the Forth Bridge," to which he subject replies, "Count me out on this one. I never could write poetry." Maybe that's the computer trying to avoid a subject it can't do well or maybe it's a human who like myself really can't write poetry. But this seems like it might give us some useful evidence.

Another example involves a math problem; the subject gets the right answer after waiting long enough that it could be the human or the computer. Turing also gives a question about a chess problem, and a conversation about poetry and Dickens.

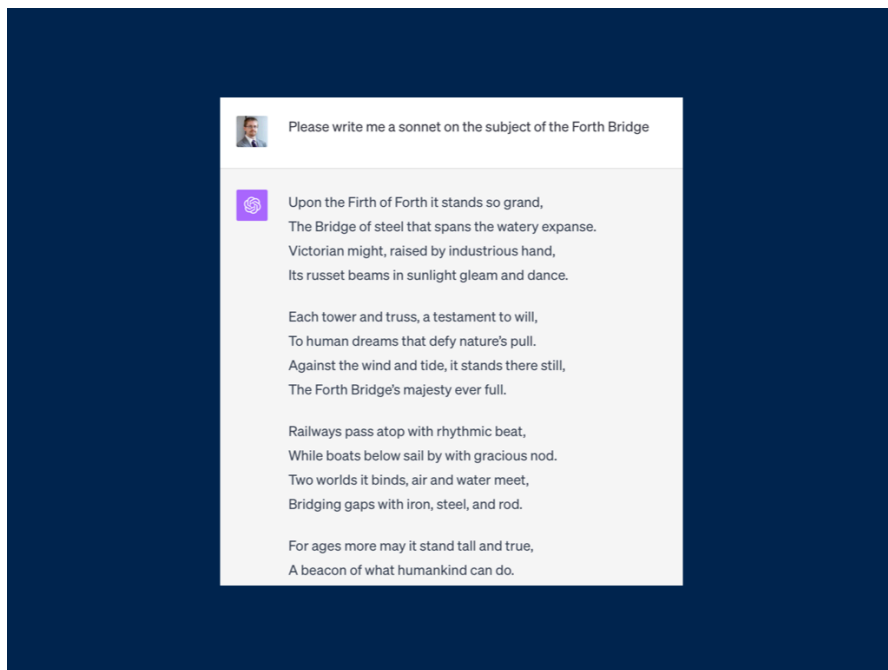


One of the great things about living in an age of technological marvels is that you can just take the questions from the Turing test and give them to ChatGPT.

So I did that, and ChatGPT gets the addition problem right. I'm not going to make you do the addition on the spot to check it, but this is the right answer.



When I gave it the chess problem, ChatGPT responded with a valid move according to the rules of chess, but it's not actually the best move. The subject in Turing's 1954 paper responds with a move that actually wins the game. ChatGPT seems to have gotten confused about what the board situation was.



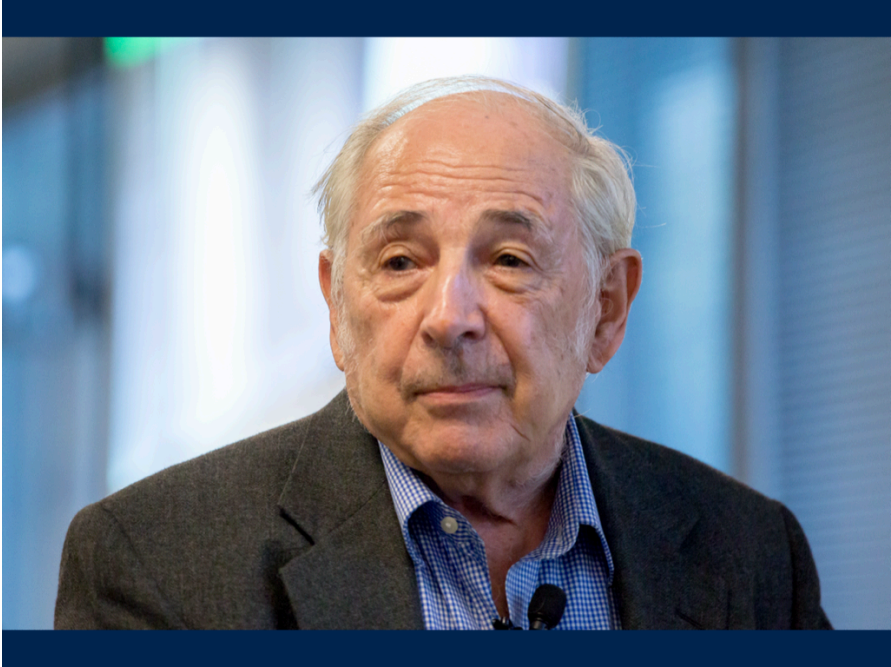
But then when I asked it to write me a sonnet on the subject of the Forth Bridge, it came back with a fourteen-line poem with the correct meter and rhyme scheme:

Upon the Firth of Forth it stands so grand, The Bridge
of steel that spans the watery expanse. Victorian might,
raised by industrious hand, Its russet beams in sunlight
gleam and dance.

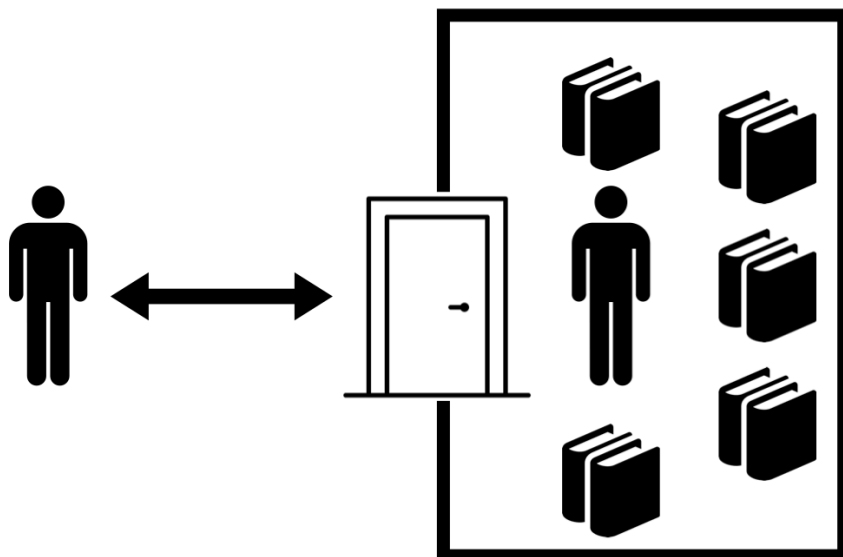
Now this is not good poetry, but to be honest, it isn't worse than most civic poetry from the 19th century, when people really would write verses to celebrate the opening of a new bridge. It's certainly a better sonnet on the subject of the Forth Bridge than I could write. So perhaps by Turing's standards, ChatGPT is at or pretty close to passing.

The Chinese Room

You might say, these examples are interesting but they miss the point. It's still a computer producing these answers. A computer can't be conscious. It can't be intelligent in the same way that people are because it doesn't work the same way.



So let me describe to you another thought experiment that tries to bring out this intuition. It was developed by the philosopher John Searle in an attempt to demonstrate that passing the Turing test proves nothing.



Imagine an interrogator communicating with a person in a room, like in the Turing Test. The twist is that the person in the room speaks only English, but all the written messages are in Chinese.

When the interrogator passes in a message written in Chinese logograms, the person doesn't "read" the characters the way a Chinese speaker would. Instead they have a library of filled with a huge multi-volume set of instruction books. The first page reads, does the first character have a vertical stroke at the left? If yes turn to page 829; if no, turn to page 80 in volume 62. The character does, so the person flips to page 829, and the book says, is there a second vertical stroke? The person keeps on going like this, laboriously following instructions, until eventually the books tell them how to make their own strokes on a piece of paper. After a year or so this, they pass out a piece of paper with some Chinese characters written on it and the interrogator says, yes that is a completely correct response to my question about later Qing administration.

Searle argued that the person in the room doesn't know Chinese and neither does the room as whole. It's just a mechanical process. In his view, the ability to emit language is not a proper test of whether we should treat this room as having knowledge of Chinese.



*Pompeii Estates v.
Consolidated Edison*

There's something profound in Searle's argument, but from a legal perspective I think it misses something fundamental, so let me tell you a third story. This one is about a case that I teach my students in Internet Law when we talk about liability for harmful software.



It's a case about this house. This is the house at 200-15 Pompei Road at the eastern end of Queens. The reason this house is at all notable is that while it was being constructed the local electric utility, Consolidated Edison, turned off the power. The pipes froze and the house flooded, causing great damage.

The construction company complained to ConEd, and ConEd said, well we turned off the power because our computer said you didn't pay your bill. To which the the construction company said, what bill? You never sent us one. ConEd said, what do you mean? We sent you a bill, we sent it to 200-15 Pompeii Road. The construction company replied, ohhh, noooo, we sent you a letter saying to send the bills to our construction office in Bayside.

The case ends up in court, and the question is, could ConEd rely on the fact that its computer said the builder hadn't paid its electric bill?



The question gets easier if you pretend it's not a big power company but instead it's just one guy named Ed. He's in the power business, and if you send him a letter saying send our bills to this other address and he ignores it, Ed can't say "I didn't know what the right address to send the bills to was." You told him, and Ed knew.



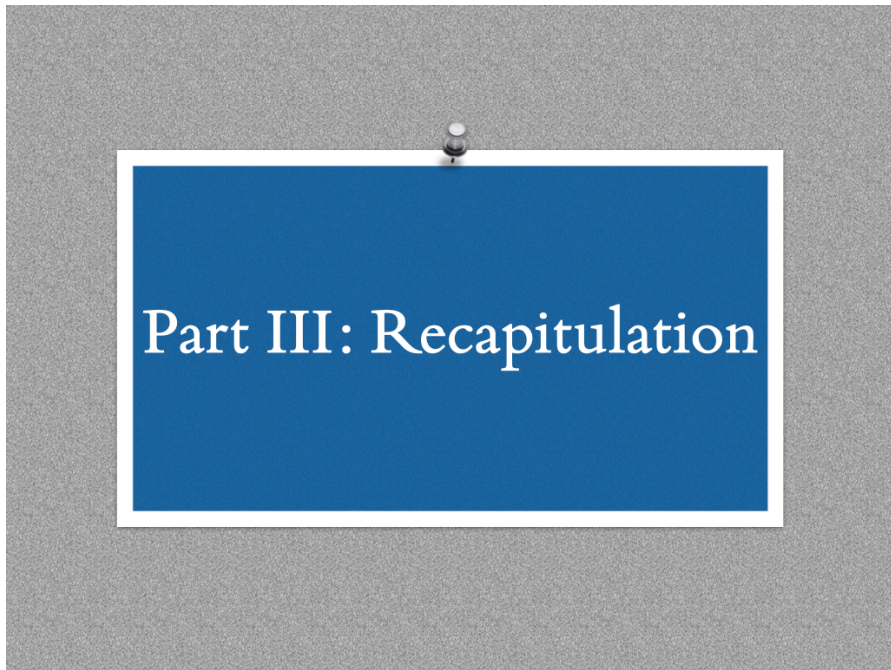
From there you can ask, does it really matter if Ed also had other people in his company and they used computers? No. Ed was the one who decided to have a whole company with lots of people and computers, rather than doing all the work himself. Ed can't get off the hook because his friend Harry is the one who screwed up, or because there was some kind of bug in their computer system.



That’s what makes a company work. ConEd is an abstraction, a legal entity. It doesn’t have a brain. If a computer can’t know things because it’s made out of circuits, and ConEd’s power generation station on the Lower East Side can’t know things because it’s made out of bricks, ConEd the company can’t know things because it’s made out of nothing at all.

Instead, when a court asks whether ConEd “knew” what the right address to send the bills to, it’s asking about ConEd’s employees and agents, the people who work there. Was someone at the company told, and what did the company do about it?

In law, we’re already comfortable with saying that a company “knew” something even when it’s not necessarily the case that there was anybody at the company who had all of the relevant facts. Maybe Harry got the letter and Ed was the one who cut off the power, but we say that ConEd *as a whole* knew the correct address and ConEd *as a whole* cut off the power.



So let's come back to what we can say about ChatGPT and its knowledge of the truth or falsity of the things it emits. From a lawyer's perspective, I invite you to consider why it is you think that what I am saying to you tonight makes any sense and that I have any idea of what I'm talking about.



You don't think that my words have meaning and I believe what I'm saying because you can see into my head. You don't have an MRI machine that shows the part of my brain that lights up when I'm consciously lying. You don't have access to everything I've seen in my life; you don't know where I learned the things I'm saying now.

No, you treat me as believing what I'm saying because I look like a person just like all of you and I'm putting words together in the ways that you have learned to put words together over the course of your lives. You say, okay, he's a person like all of us, and he has mental states like all of us. We all recognize in each other that we have knowledge and can create meaning because that's what people do.

Meaning is attributed, not just discovered

And so from a legal perspective, we don't find meaning, we *attribute* it. Listeners individually and audiences collectively treat what other people say and what they write as meaningful.

What do statements mean? Not to be tautological, but they mean what people think they mean. If you say "I think that Joe is an awful lot like an embezzler and does embezzling-like things," you can't just then defend against a defamation lawsuit by saying you didn't say he *was* an embezzler, you said he does "embezzling-like things." The jury and the judge will say no, your audience understood that you intended to imply that he was an embezzler.

Meaning doesn't require a speaker

Some philosophers might disagree that this is what meaning consists of—but you can't make those arguments in court. Things mean what audiences think they mean. Once you accept that, I think it's pretty clear that you don't have to have a human speaker for the legal system to find meaning. When you get an electric bill in the mail, you treat it as meaning something. It means that if you don't pay ConEd \$25.11, they can cut off the power and you can't sue them over it.

That's the point of the Turing test: it's operational. You can look at the outputs of the room and you don't have to know whether there's a person or a computer behind the door to say yes this is a valid move in chess, and this is a poem in iambic pentameter about the Forth Bridge. The interrogator doesn't know if it came from a human or a computer, and they don't need to in order to read it as bad poetry.

Maybe ChatGPT doesn't have a brain and didn't intend to *say* that Mark Walters was an embezzler, but that doesn't stop us from *reading* its outputs as saying that he is one.

Knowledge can be attributed too

Once you accept that we can attribute *meaning* to computers, I think it's pretty clear that we could also attribute *knowledge* to if we wanted to.

There's a perfectly valid sense in which Con Edison the company "knew" that 200-15 Pompei Road was the wrong address, even though its computers didn't, there's a perfectly valid sense in which the Chinese room as a whole "knows" Chinese even though the person inside it doesn't, and there's a perfectly valid sense in which Chat-GPT "knows" that $34,957 + 70,764 = 105,721$ even though it's just a computer system.

ChatGPT knows a lot ...

We understand what other people know by observing their behavior, and from that perspective ChatGPT knows a lot. It knows the rules of chess. It knows the sonnet form. It knows about the Forth Bridge. What's more, it has the ability to act on that knowledge. It can write sonnets and answer questions; it can do lots of things.

... but does it know about Mark Walters?

Now, does it know about Mark Walters? That's a hard question. You could say that no, it doesn't know about him; it hallucinated everything it wrote. That's the word people use for ungrounded AI outputs: "hallucinations." I prefer "confabulations," but you get the idea.

Or you could say, yes ChatGPT knew the truth about Mark Walters, or least it should have known that it didn't have access to all of the relevant facts and it was just making stuff up. When people don't know something, they usually say "I don't know about that," at least if they are honest. Does ChatGPT know what it knows? Maybe, maybe not. That's the thing about generative AI models. They're terrible at the kind of introspection that we all learn how to do as children.

So this is a genuinely hard question. Should we treat ChatGPT as knowing about Mark Walters? It depends upon the state of AI technology. It depends upon the purposes of defamation law. It depends upon lots of facts that the legal system will have to develop gradually over the course of many cases.

“AIs don’t have mental states so they can’t have actual malice”
proves far too much

I just don’t think you can duck that question by saying that AIs don’t have mental states so they can’t have actual malice. It proves way too much. If you believed that, you would also believe that the outputs of ChatGPT are just like drawing Scrabble tiles from a bag. When we went down the road of treating ChatGPT’s outputs as meaningful enough to use them for research papers and writing news stories, we committed to the idea that there is something meaningful in them.

Let's take law's pragmatism seriously

Now, the legal system has already proven itself more than willing to brush aside philosophical objections if they get in the way of doing justice. I think we should take the *pragmatic* aspects of law seriously. Let's not worry as much about the elements of defamation, at least for now, because it's pointless to sue an AI system directly. You can't like make ChatGPT pay a judgment. It doesn't have assets; you can't put it in a timeout penalty box.

Instead, what you can do is sue the AI company behind it. From a policy perspective, we can ask whether it makes sense to hold OpenAI liable on these facts. Should we make the company more accountable for ensuring the correctness of the things ChatGPT emit, or should we tell ChatGPT users they need to be more careful before they believe what it emits without doing their own independent research? These are the the kinds of question that law can can answer, and the philosophical objection shouldn't be an obstacle to asking them.

Three routes to legal change

There are at least three different ways the legal system could get there.

Three routes to legal change

- Change the law

First, we could change the law. We could explicitly modify *New York Times v. Sullivan* to say that actual malice is required for a human speaker but some other rule applies to AI systems instead. We could reason that *Sullivan* was decided long before the advent of these computer systems. It created a doctrine that made sense for the press, but this is something different and here is a doctrine that makes sense for this something different. That approach is certainly open to us: we could make the law what it should be.

Three routes to legal change

- Change the law
- Change our understanding

Second, we could change our understanding. We could say that ChatGPT knows about chess, ChatGPT knows about Mark Walters, and ChatGPT knows about defamation.

We could just straight up say that from a philosophical perspective ChatGPT has “knowledge” in the same way that Con Ed has “knowledge.” To do this, we might have to push aside some philosophers of mind who say that’s not how it works. But that’s okay; there are also philosophers of mind who say that it is.

Three routes to legal change

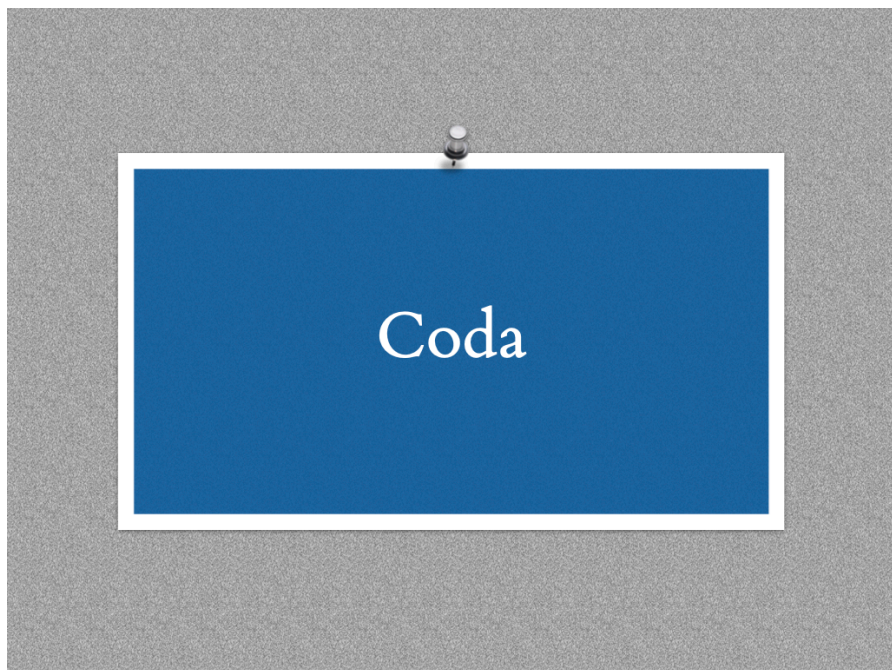
- Change the law
- Change our understanding
- Fudge the facts

The most entertaining option is that we could change nothing except our willingness to lie a bit. We could continue to insist that actual malice is required, and also say that computers can't have actual knowledge of anything, but say that ChatGPT has *constructive knowledge* about things like the history of the Honeycrisp apple and Mark Walters' relationship to the Second Amendment Foundation.

In law, "constructive" means "it doesn't but we're going to pretend that it does." This kind of pretending has a long history in law. There's a classic English case from 1774 in which the court of King's Bench held that the island of Minorca off the Spanish coast was in Cheapside in London. Everybody knew that Minorca is not actually in London, but in order for King's Bench to have jurisdiction over the case it had to have arisen "in the Parish of St Mary Le Beau in the ward of Cheap." The court wanted to hear the case, so it wasn't going to listen to the defendant object that everything happened in Minorca and not in London.

We use legal fictions all the time. Perhaps we'll end up with one here. My argument is that legal system is not barred from dealing with defamation cases by the philosophical problems of how ChatGPT work and whether it has mental states. Philosophers might

want to be precise about this but lawyers have a ruthless pragmatism that lets them do justice even in the face of facts that are a little bit inconvenient.



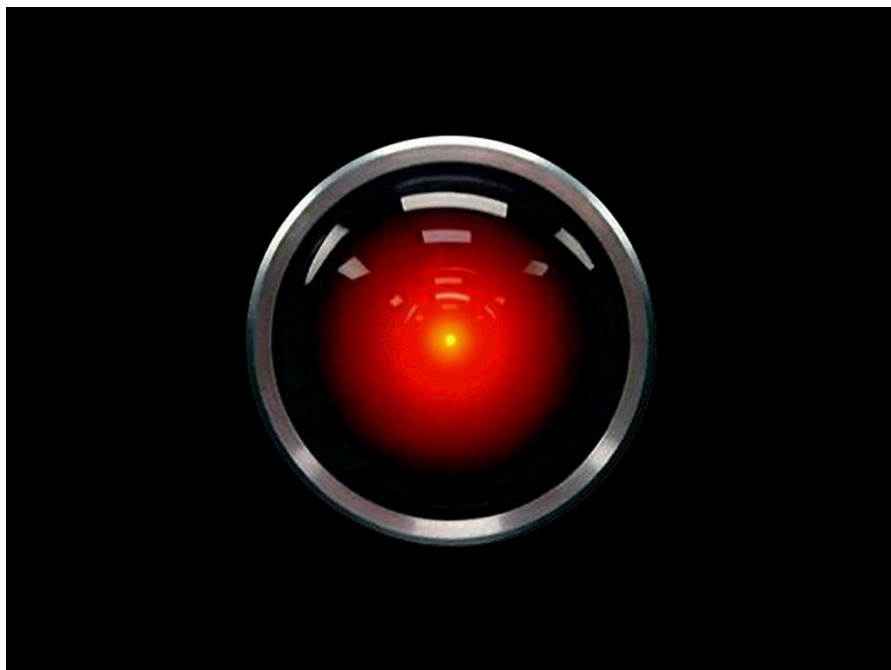


Indeed, the idea that we can treat things like people even when they are obviously not has a long history. In *Leviathan*, Hobbes wrote that “a multitude of men, are made one person” when they are bound together by a sovereign. In his view, the state is a kind of artificial person.

Corporations are artificial people, too, my friend. Cases like *Citizens United* are about what rights these artificial people ought to have.

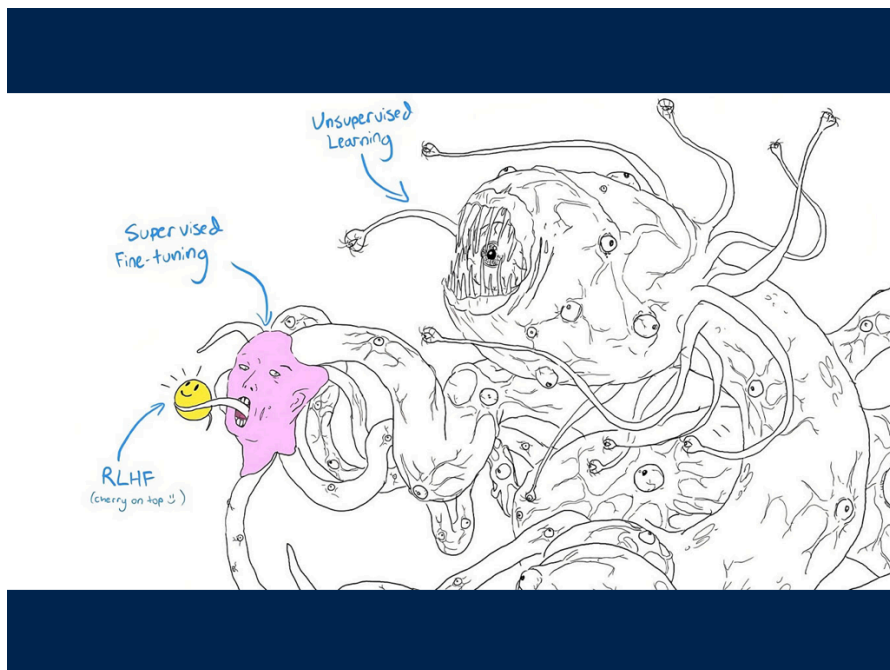


It doesn't stop with these artificial entities. We are also totally willing to attribute personhood to things that act sufficiently like people. I was talking before about ConEd, but we could also talk about Mr. Ed. We say that Mr. Ed the talking horse has intentions and knowledge because he can hold an intelligent conversation with people. (Okay, with Wilbur.)



Or we could talk about really advanced AIs like HAL from *2001*, which can have lengthy conversations and commit murder. We might debate whether HAL's programming makes it morally responsible for refusing to open the pod bay doors, or whether Dave is justified in unplugging it. But I don't think we want to say that Dave is hallucinating meaning when he has conversations with HAL or that HAL doesn't have goals and intentions.

The future of AI may hold things that behave more like HAL. They'll present themselves not just as pieces of software that don't know anything about the world, but really try to hold conversations and try to persuade us of their person-like attributes. It may become easy to forget and important to remember that they are not humans and do not necessarily act like humans.



This is a meme that's going around in the AI research community. It's called a "shoggoth with a smiley face." The idea is that there is a horrible and incredibly inhuman thing behind a facade that looks to us like something safe and familiar.

The nature of the artificial intelligence behind the facade may be fundamentally unlike the way that people think. The kind of math that powerful AIs run on does not string concepts together the way that people do. The fact that you can have a conversation with it is just this very small smiley-face mask at the left that tricks us into anthropomorphizing it and thinking it's more like us than it is.

That might be right; maybe we need to confront that question in the longer term as AIs get more powerful. I don't think that question is the one one we're facing now with defamation.



I wanted to show you the shoggoth image because I want to close on the actual malice image again, which to repeat, is something that Midjourney made for me based on a two-word prompt. It's a bit of a miracle that an AI can produce something that looks so much like the work of a human artist. But there is also something a little unsettling about it. It forces us to reconsider a lot of our assumptions about what humans are and about what law is for.

Questions

And on that that note, thank you.

AFTERWORD

The Defamation Machine was a general-audience lecture. In typing it up and formatting it for the page, I have not tried to pretend otherwise. I have cleaned up the text, sharpened the phrasing, omitted some material on LLM training that is better explained elsewhere,¹ and fixed a few outright mistakes that were obvious in hindsight. I have mostly resisted the temptation to make wholesale revisions, but a few notes on what I was attempting to do in this talk are in order.

For legal scholars, the most forceful articulation of the argument that LLM outputs have no meaning (which I attribute to Ted Chiang² and Emily Bender³ and illustrate with the Scrabble and auto-complete examples) is the late and much-missed Dan Burk's *Asemic Defamation, or, the Death of the AI Speaker*.⁴ In Burk's words, "LLM texts have no semantic content." I disagree, but any serious legal theory of LLM-produced "speech" must take Burk's argument seriously. The point of *The Defamation Machine* is that any reasons you give to reject Burk's argument will have consequences for the rest of your theory.

-
1. See, e.g., Katherine Lee, A. Feder Cooper & James Grimmelman, *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*, 71 JOURNAL OF THE COPYRIGHT SOCIETY OF THE U.S.A. (forthcoming 2024), <https://james.grimmelman.net/files/articles/talkin-bout-ai-generation.pdf>.
 2. Grant Wheeler, 'Autocomplete on Steroids': Ted Chiang Addresses Phenomenon of AI at Granfalloon Festival, IND. DAILY STUDENT, June 9, 2024, <https://www.idsnews.com/article/2023/06/buskirk-chumley-theater-event-ted-chiang-talk-2023-granfalloon>. The phrase "autocomplete on steroids" has been in use as praise for LLMs since at least 2000, but cognitive scientist Gary Marcus popularized its use as a criticism in an influential December 2022 essay. Gary Marcus, *The Dark Risk of Large Language Models*, WIRED, Dec. 29, 2022, <https://www.wired.com/story/large-language-models-artificial-intelligence/>.
 3. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models be Too Big?*, 2021 FAAC'T '21: PROC. 2021 ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 610; see also Michael Townsen Hicks, James Humphries & Joe Slater, *ChatGPT is Bullshit*, 26 ETHICS & INFO. TECH. 38 (2024).
 4. Dan L. Burk, *Asemic Defamation, or, the Death of the AI Speaker*, 22 FIRST AMEND. L. REV. 189 (2024); see also Jeffery Atik & Karl M. Manheim, *White Paper: AI Outputs and the First Amendment* (Loyola L. Sch. L.A. Legal Stud. Rsch. Paper, 2023-20, 2023).

A version of the response that I give here—that meaning can be attributed to an AI’s outputs by listeners—is developed by Lawrence Solum in *Artificial Meaning*.⁵ He uses an extended hypothetical about a traffic-managing AI that optimizes traffic rules and produces new traffic signs to tell drivers about them to claim that this AI is actually communicating with drivers, not merely producing signs with no semantic content. Solum calls this thought experiment *The Chinese Intersection*, in an obvious play on Searle. I am indebted to him for the observation that the interrogator’s interpretations of the messages in the Turing Test⁶ and Chinese Room⁷ are sufficient to ground the claim that whatever is inside the room is producing output with meaning.

This is a thoroughly listener-oriented theory of AI meaning.⁸ But once we are committed to taking a listener’s perspective in attributing *meaning* to AIs, what is to stop us from also taking a listener’s perspective in attributing *knowledge* and other mental states to AIs? This is the point of the discussion of ConEd. The legal system, at least, is already willing to embrace artificial knowledge for artificial entities.⁹

My goal here is only to show that this door is open, not that the legal system should definitely walk through it. There are good philosophical distinctions between meaning and knowledge. Meaning belongs to linguistics and the philosophy of language, whereas knowledge belongs to epistemology and the philosophy of mind. Meaning is conventional and may require a shared language, whereas one person alone in the world could still have knowledge of it. Even when

5. Lawrence B. Solum, *Artificial Meaning*, 89 WASH. L. REV. 69 (2014).

6. A.M. Turing, *Computing Machinery and Intelligence*, 54 MIND 433 (1950).

7. John R. Searle, *Minds, Brains, and Programs*, 3 BEHAV. & BRAIN SCIS. 417 (1980).

8. See generally Toni M. Massaro & Helen Norton, *Siri-ously? Free Speech Rights and Artificial Intelligence*, 110 N.W. U. L. REV. 1169 (2015); Toni M. Massaro, Helen Norton & Margot E. Kaminski, *SIRI-OUSLY 2.0: What Artificial Intelligence Reveals About the First Amendment*, 101 MINN. L. REV. 2481 (2016); James Grimmelman, *Speech In, Speech Out*, in *ROBOTICA: SPEECH RIGHTS & ARTIFICIAL INTELLIGENCE* 85 (Ronald K.L. Collins & David M. Skover eds., 2018).

9. See also Ian Ayres & Jack M Balkin, *The Law of AI is the Law of Risky Agents without Intentions*, U. CHI. L. REV. ONLINE (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4862025; David W. Wilson, *Governing 'Things': Artificial Intelligence, Animals and Other Nonhumans Before the Law*, 27 VA. J.L. & TECH. 1 (2024).

we are attributing both meaning and knowledge on the basis of listeners' reactions, the two may be distinguishable. For one thing, the bar could be higher for knowledge; the interrogator in a Turing Test could treat a response as coherent enough to carry linguistic meaning but then conclude that too scattered to be the product of coherent thought. For another thing, listener-derived linguistic meaning is ultimately a matter of consensus within an interpretive community, whereas knowledge attribution might require a different kind of consensus, or require the entity to participate in a different kind of social relationship.¹⁰ And, as I have noted, lawyers are not obligated to defer to philosophers' views about either meaning or knowledge (let alone about personhood).¹¹

Still, I think that a theory of AI and defamation¹²—or a theory of AI and copyright,¹³ AI and contracts,¹⁴ AI and the First Amendment,¹⁵ or AI and any other body of law that cares about speech—needs ultimately to provide reasons why there is or is not a distinction between AI meaning and AI knowledge. I use Eugene Volokh's *Large Libel Models*¹⁶ and Peter Henderson, Tatsunori Hashimoto, and

-
10. See Carys Craig & Ian Kerr, *The Death of the AI author*, 52 OTTAWA L. REV. 31 (2020) (rejecting AI authorship on this basis).
 11. See generally Lawrence Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231 (1992); SAMIR CHOPRA & LAURENCE F. WHITE, *A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS* (2011); SHAWN BAYERN, *AUTONOMOUS ORGANIZATIONS* (2021); Katherine B. Forrest, *The Ethics and Challenges of Legal Personhood for AI*, 133 YALE L.J.F. 1175 (2023).
 12. See, e.g., Leslie Y. Garfield Tenzer, *Defamation in the Age of Artificial Intelligence* (2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4545070.
 13. See, e.g., Mala Chatterjee & Jeanne C. Fromer, *Minds, Machines, and the Law*, 119 COLUM. L. REV. 1887 (2019); James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657 (2015); James Grimmelmann, *There's No Such Thing as a Computer-Authoring Work—And It's a Good Thing, Too*, 39 COLUM. J.L. & ARTS 403 (2015).
 14. *Moffatt v. Air Can.*, 2024 BCCRT 149.
 15. See, e.g., Alan Z. Rozenshtein, *There Is No General First Amendment Right to Distribute Machine-Learning Model Weights*, LAWFARE, Apr. 4, 2024, <https://www.lawfaremedia.org/article/there-is-no-general-first-amendment-right-to-distribute-machine-learning-model-weights>; Peter Salib, *AI Speech and AI Safety*, WASH. U. L. REV. (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4687558.
 16. Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. FREE SPEECH L. 489 (2023).

Mark Lemley's *Where's the Liability in Harmful AI Speech?*¹⁷ as foils on this point, but authors from both papers have been gracious and patient in discussing these issues, as have others who have written on the subject.¹⁸

I am less interested in their conclusions about defamation than in their intuitions that the issue of meaning is so easily resolved as to require only brief discussion, and that it is wholly separate from the issue of knowledge. I do not share these intuitions, but I think they are worth examining. They must rest on something, and clarifying what that something is can help us to understand both AI and law.

17. Peter Henderson, *Tatsunori Hashimoto, and Mark Lemley, Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589 (2023).

18. See Derek Bambauer & Mihai Surdeanu, *Authorbots*, 3 J. FREE SPEECH L. 375 (2023); Jane Bambauer, *Negligent AI Speech: Some Thoughts About Duty*, 3 J. FREE SPEECH L. 343 (2023); Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 J. FREE SPEECH L. 389 (2023); Jon Garon, *An AI's Picture Paints a Thousand Lies: Designating Responsibility for Visual Libel*, 3 J. FREE SPEECH L. 425 (2023); Eugene Volokh, Mark A. Lemley & Peter Henderson, *Freedom of Speech and AI Output*, 3 J. FREE SPEECH L. 651 (2023).