Chatbots and Speech Act Responsibility Gaps

Guido Löhr

Vrije Universiteit Amsterdam

loehrg@icloud.com

g.lohr@vu.nl

To appear in *Communicating with AI: Philosophical Perspectives,* edited by Rachel Sterken and Herman Cappelen. OUP.

**Abstract**

I argue that chatbots create a peculiar new kind of responsibility gap, which I call the “speech act responsibility gap”. Unlike the responsibility gaps commonly discussed in the context of self-driving cars and autonomous weapons, speech act responsibility gaps arise from the fact that paradigmatic speech acts like assertions (statements), promises or orders always generate linguistic commitments and entitlements. Unlike more familiar kinds of responsibility gaps, speech act responsibility gaps are inherently interpersonal and directed. I first argue that currently dominant treatments of chatbot speech acts as proxy agents cannot bridge these gaps. I also discuss why current arguments against the existence of responsibility gaps don’t apply in the case of chatbots. Instead, responsibility appears to be best attributed to the chatbot itself. However, this poses a dilemma. Either these machines don’t speak (we are fundamentally mistaken about their output), or we need to engineer (broaden) our notion of responsibility.

**Keywords**

Speech Acts; Responsibility Gaps; AI; Large Language Models; Chatbots; Conceptual Disruption; Conceptual Engineering; AI Assertion; Assertions; Entitlements

**Conflict of interest**

I have no conflict of interest to declare.

**Introduction**

Matthias (2004) directed our attention to the problem of “responsibility gaps” with “learning automata”. Such gaps arise in light of intelligent learning systems that do not just execute rule-based instructions but adjust their neural networks automatically to adapt to the respective context. This form of “machine learning” gives humans less control over the AI’s output or “behavior”. Moreover, large artificial neural networks are black boxes whose inner workings need to be studied in a similar fashion as we study biological neural networks (Zednik, 2021; Buckner, 2019). This makes AI often not only difficult to control but also difficult to predict (Boge, 2022; Sullivan, 2022).

The lack of control and predictability of cutting-edge machine learning algorithms "disrupt" (Löhr, 2022; Sterken, 2020) our ordinary concept of responsibility, whose attribution seems to require both, i.e., a certain degree of predictability and control (Coeckelbergh, 2020; Nyholm, 2023).[[1]](#footnote-1) It is difficult to hold another person responsible for an event that they were not in control of and that they could not have foreseen and therefore not be negligent about (Veluwenkamp & Hindriks, 2024). Lack of control without negligence is normally considered a justified excuse for wrongdoing (Tigard, 2021; Strawson, 1962/2008). The worry is that the introduction of learning systems would, therefore, diminish the possibility of holding people accountable (Tollon, 2023).

The learning systems that have received the most attention in the responsibility gap literature are self-driving cars and autonomous weapon systems (see Nyholm, 2023 for a review). If self-driving cars have the potential to cause harm that nobody can be held responsible for, then this is a *pro tanto* reason against introducing this technology (Nyholm, 2018; List, 2021). A similar argument has been made against the introduction of autonomous weapon systems (Sparrow, 2007; Swoboda, 2018). It seems to speak against such systems that they may generate atrocities that were difficult or impossible to predict and control and that, therefore, nobody can rightly be held accountable for (cf., Oimann, 2023). However, worries have surfaced about whether such cases are really responsibility gaps rather than, e.g., simply accidents (Köhler et al., 2017; Tigard, 2021).

In this paper, I discuss a technology that has not been discussed in the context of responsibility gaps but for which gaps arise in a special, more convincing, and even peculiar manner. The technology that I focus on here is AI-powered chatbots, i.e., applications that rely on large language models (LLMs) that specialize in natural human language.

The problem I will discuss is this: If the output of chatbots is supposed to emulate communication (speech acts), then we cannot help but use normatively rich terms that entail the ascription of assertory commitments and entitlements as well as promissory obligations and rights to describe what sophisticated chatbots are “doing”. This, I argue, gives rise to a new kind of responsibility gap that is substantially different from the responsibility gaps that are currently the focus of the debate. The gaps, I am talking about are essentially directed and interpersonal. This means that the entitlements and commitments in speech are always directed at the respective speakers.

How can we best think about speech act responsibility gaps? Currently, most researchers who have worked on chatbots and speech acts recognize an epistemic responsibility gap but claim to have filled it by positing proxy agents to ascribe responsibility not to chatbots themselves but to their owners (e.g., Nickel, 2013; Green & Michel, 2022; Butlin & Viebahn, forthcoming). However, I will show that we cannot help but attribute them to the chatbots themselves if we want to make sense of our interaction with them.

This creates a dilemma. Either we need to broaden (i.e., a form of conceptual engineering, Cappelen, 2018; Koch et al., 2023; Löhr, 2024) our notion of responsibility (cf., Tigard, 2021, pp. 442–443 for a similar suggestion with other AI systems). Or we need to conclude that AI utterances do not amount to real speech, which means that we have no language to make sense of what they are doing and how we should engage with them. The former, I argue, is the more attractive horn of the dilemma.

**2 Responsibility gaps**

Roughly speaking, responsibility gaps occur in situations where we would normally look for someone to hold responsible (and where the concept of responsibility would normally have an important societal function), but where no candidate could appropriately be held accountable (cf., Köhler, et al., 2018, p. 54). In most cases, this is because two common conditions for attributing moral responsibility are not met (Nyholm, 2023; Matthias, 2004). First, no human agent is in control of the event and, second, no human agent could have predicted it.

Note that such gaps are not merely epistemic (someone is in fact responsible, but we simply do not know who), but metaphysical (Oimann, 2023). Either nobody can be held accountable or the conditions for being held accountable still need to be introduced in society, say, by introducing a new law. In the latter case, we would say that the responsibility gap was “bridged” or “filled” by this law, thereby acknowledging that it existed. We would also not speak of responsibility gaps if the action in question could be excused (cf., Tollon, 2023). In such a case, we would say that the person was causally responsible but not morally responsible because she had no control over the relevant object – say, she was drunk or otherwise incapacitated. A lively debate in the philosophy of AI has developed on how to identify responsibility gaps and how to bridge them (Nyholm, 2023; Oimann, 2023 for reviews).

According to Matthias (2004) and others, responsibility gaps primarily occur with self-learning algorithms. The paradigmatic example of a metaphysical responsibility gap is an accident with a self-driving car (Nyholm, 2018a). In 2018, Elaine Herzberg was the first person to die in a collision with a self-driving car. This case has generated a heated debate. Normally, in such a situation, we are looking for someone that we can hold accountable. We do so because attributing accountability plays several important roles in society from identifying those who can compensate the victim to enforcing and establishing societal norms.

However, because a self-learning algorithm was involved in the death, many philosophers saw this case as an instance of a responsibility gap (e.g., Nyholm, 2023). Others rejected this notion (e.g., Tigard, 2021). Despite this disagreement, it is widely agreed that this is a situation where we normally look for a responsible agent, but where neither the typical culprit – the human driver – nor the company UBER could have anticipated and, in an obvious way, avoided the accident – both of which are key standard conditions for responsibility. (Eventually, the human driver pleaded guilty to endangerment and was sentenced to three years of supervised probation.)

Another example of a supposed responsibility gap comes from Matthias's (2004) original set of examples – an adaptive elevator system that is in operation in many high-rise office buildings:

In high-rise office buildings, adaptive elevator systems (...) analyse traffic patterns, typically using artificial neural networks and reinforcement learning algorithms, and they try to minimise waiting and transportation time for the users of the elevator. Now let us suppose, that such a learning, adaptive elevator leaves an important executive waiting for half an hour in the 34th floor, so that he cannot attend a business meeting at which he is expected. A considerable financial damage is caused. Who can be held responsible? The manufacturer can deny responsibility because the elevator, being able to learn, had changed the parameters of its program during the course of its operation, so as to better adapt to the traffic patterns in the building. Because of this, it was no longer possible for the manufacturer to predict or control the specific behaviour of the elevator in a given situation. And since the manufacturer could not have predicted or averted the undesirable outcome in this case, he cannot justly be held responsible for this specific behaviour of the machine. He might nevertheless be held responsible for selling elevators of this type, if the dangers of using them are not advertised properly in the operating manual. But this is a different point, and, as we will see later on, there might not have been any choice of an alternate, ‘safe’ system at all.

Like the Uber case, the elevator case also involves a learning mechanism that reduces the control and predictability of the engineers and the owner. What distinguishes the elevator case from the Uber case is that while both are learning systems, the elevator worked properly, at least as advertised (though with an unintended side-effect), while the Uber car generated an unwanted event due to a shortcoming of the system. Thus, in both cases, one could object that no gap exists because we would not expect anyone to be held responsible in this case (e.g., Tigard, 2021; Köhler, 2017). Sometimes things simply don’t happen as planned, the company could have tested more, or the driver should paid more attention.

In the next section, I show that a responsibility gap occurs in a different but arguably much more obvious way in the case of chatbots. I call this distinct kind of responsibility gap – a "speech act responsibility gap". I argue that it occurs even if critical voices like Tigard (2021) are right about the lack of more familiar kinds of responsibility gaps. Thus, I argue that speech act responsibility gaps are clearer and equally pressing metaphysical gaps to grapple with, which, too, arise with self-learning systems where human agents are not fully in control. At the same time, I show that speech act responsibility gaps are quite different from the standard notion of a responsibility gap.

As we will see, speech act responsibility gaps and more traditional responsibility gaps can occur simultaneously and are interdependent in interesting ways. This fact, combined with the prediction that chatbots will soon play very central roles in our lives, renders the notion of a speech act responsibility gap interesting for philosophers of technology who are interested in the ethical aspects of chatbots (Kasirzadeh, A., & Gabriel, I. (2023; Nyholm, 2023; Dennis, 2022; Krueger & Osler, 2022). It will also interest philosophers of language who might be more interested in the question of whether chatbots can truly speak (e.g., van Woudenberg, et al., 2024; Viebahn & Butlin, forthcoming; Green, M., & Michel, 2022; Mallory, 2023 or Nickel, 2013).

**3 Speech act responsibility gaps**

Meet Alexa 2.0, a hypothetical future digital artifact that is a mix of chatbot and personal assistant (from now on I will omit the “2.0”). According to the (hypothetical) manual:

Alexa will be your main user interface (similar to Samantha in the movie *Her*). Through Alexa, you can control all your interconnected devices. You can tell it to turn on the oven, search for recipes, and control the kitchen machine to cook it. Of course, besides being an interface, Alexa is also your (possibly main) companion (think about how much more time you spend with your current operating system than with your friends). You will have great conversations with Alexa who is available for a chat and support 24/7 (without ever getting tired of your stories and ticks).

Moreover, the manual makes the following clear from the start:

Don’t expect it to get everything right immediately! Alexa is a learning system. It is designed to adapt to your personality, interests, values, and needs. Getting to know you will take a while for Alexa, perhaps even quite some time. But after a year or so, Alexa will have adapted to your preferences and interests completely. If you are not satisfied with the product after a year or so, you can safely return Alexa, no questions asked. However, you need to keep it and use it for at least a year for the warranty to be redeemable. By that time, we have collected so much data from you that we don’t mind if you return it.

The technology I am describing – the kind of neural network that underlies Alexa 2.0 – is not science fiction. In part, it is already available. For example, the app “Replika”, which uses the language model GPT, already provides a similar product for "free". Your personal Replika remembers what you have been talking about and keeps reminding you that it is still learning, that negative feedback is appreciated, and that it will help the algorithm adapt to you. Many other chat-based apps learn from your input whether it is a mere conversational app or one of the many psychotherapy apps on the market (“Better Help”, for example). They all adapt to you by adjusting their networks according to your input.

None of these apps can control your kitchen devices and conversations with them are often a bit frustrating. However, it is clear that something like Alexa 2.0 is the aim of many tech companies (cf., the company Rabbit and their chatbot operating system), and we have no reason to believe that it won’t become a reality relatively soon, unlike wild claims about singularity or AI consciousness. The latest versions of ChatGPT are already linked to other apps like Dall-E and can create a picture or generate a spreadsheet using a simple “request”. After the hype around ChatGPT in 2023, it is hardly news that conversational AI has made immense progress. Once connected to all our other internet devices, it has the potential to become *the* interface that everyone (whether old, young, educated, or illiterate) can and probably will use (Dennis, 2022; Löhr & Dennis, 2025).

The relevant output of chatbots for this paper is natural language speech. While AI-powered self-driving cars transport you from one place to another, while autonomous weapons expel bullets without the explicit instructions of humans, chatbots (arguably as we will see) are designed to perform (or simulate) what linguistics and philosophers call “speech acts”. Speech acts are actions with a speech component (Searle, 1969; Austin, 1962). While we sometimes separate actions from speech (in the expression “all talk and no action”), most of our daily actions in fact involve speech. In other words, most of our daily actions are probably linguistic.

Speech acts include mundane things like statements (also called assertions): “Today is sunny”, promises: “I promise to help you with your homework”, requests: “Can I have a drink?” or orders: “Clean the bathroom!”. Speech acts can also be declarations, e.g., when you baptize your child by uttering “I call you Anne” at the right moment. Speech acts are "things we do" (acions), by using speech (including text). This does not mean that speech acts are not efficacious. On the contrary. Promises create obligations and, according to Searle (2010), they even create our social institutions and thereby much of our social world. Even presidents are created by declarations, namely speech acts of the form “I hereby declare you president of the chess club”.

Like most (if not all) actions, speech acts are normatively charged. When ordering a coffee, you become responsible for paying for the coffee if it is delivered to you. When you tell your friend that the sun is shining, this commits you to the fact that the sun is really shining (Shapiro, 2020; Brandom, 1983). If the sun is not shining, your friend is owed an apology or an excuse. In other words, just like many actions in the public sphere, speech actions normally have normative consequences (cf., Kasirzadeh & Gabriel, 2023).

Interestingly, the type of accountability generated by many speech acts is of a slightly different kind compared to the accountability discussed in the case of self-driving cars and autonomous weapons. Obligations created via speech acts are inherently interpersonal or “directed” (cf., Wallace, 2019; Darwall, 2013; Gilbert, 2018; May, 2015). When making an assertion or promise, we always owe specific people to speak the truth or to act on the promise. If I tell you that the bank is open on Sundays, I owe it *to you* that what I said is true and that I have a good reason for thinking and communicating it. This accountability of the hearer is special in the sense that it is directed at *certain* individuals but not others. The same goes for promises. If I promise to mow your lawn, *you* can hold me accountable for failing to do so in a special "personal" way that isn't on anyone's behalf but you (Strawson, 1962/2008; Darwall, 2013).

The point of departure of this paper is that all the interactions just described, which use natural human language as the primary medium, can currently only be described with the only words available to us, which developed to express human speech acts. We have no other language (no other concepts) but our current ones (assertion, promises, order, question, request) to describe what is going on when we talk with chatbots. To illustrate this, I take the following to be typical interactions with Alexa 2.0 or even current versions of ChatGPT:

**So-called “constatives” (e.g., assertions, announcements)**

Alexa: “The movie starts at 8 pm”

User: “Thanks, Alexa!”

User: **“**Alexa, the movie did not start at 8 pm, this was incorrect.”

Alexa: “I apologize!”

**So-called “directives” (e.g., orders, requests, instructions)**

User: “Alexa, rent the latest Terminator movie for me please.”

Alexa: “Ok, sure!”

User: “Alexa, find a recipe for a nice Lasagna, please. I have 5 friends over tonight.” – Alexa: “Ok, sure!”

**So-called “commissives” (e.g., promises, agreements, pledges)**

User: “The movie started at 8 pm in Hong Kong, but I am back in Amsterdam!”

Alexa: “I promise to be more careful about the time in the future.”

User: “The recipe is for 2 people! And you should know I am a vegan.”

Alexa: “I promise that I will find a better recipe, give me a second.”

As mentioned, like most other types of social action (buying coffee or taking a walk with a friend in a crowded area), speech acts normally have normative consequences. In fact, they are difficult to make sense of *without* ascribing something like an accountability relation or entitlement relation to the speakers. It is not easy to think of what an assertion could be without ascribing some form of commitment to the truth to the speaker in the sense that this person would be held accountable if the assertion turned out to be false (cf., Brandom, 1993; Shapiro, 2020; Marsili & Green, 2021).

This applies equally to assertions made by the chatbot and assertions made by the user toward the chatbot.

A natural conversation between a user and a chatbot will likely involve not just instructions coming from the human agents but negotiating these commitments and entitlements. This lies in the nature of a conversation (Geurts, 2019). Much of our conversations are spent on giving reasons for our assertions or challenging them (Brandom, 1993). In conversations, we thereby create a kind of “conversational score” (Lewis, 1979), i.e., some kind of recording or memory about who said what and when. Every unchallenged assertion will become part of the common ground of the speakers – a body of information that we are entitled to presuppose when communicating (Stalnaker, 2002).

What goes for assertions, also goes for requests, orders (directives), and promises (commissives). If successful, a request normally commits the hearer to either follow the request or give a justification for the rejection of the request (Darwall, 2013; Searle, 1969; Austin, 1961). Orders are such that they give the other person an obligation to act according to the order (if it is a legitimate order) without requiring an extra justification. It seems to be the nature of an order that no justification needs to be given – otherwise, it would be a request (cf., again, Darwall, 2013).

Promises, too, are difficult to make sense of without ascribing obligations and entitlements to the speakers (Habib, 2022). What could a promise be if not an action that generates an obligation (Gilbert, 2018)?[[2]](#footnote-2) At least, promises generally or mostly generate obligations even if it is not essential for them. Still, I assume here that we have no other way but to interpret a lot of the utterances of chatbots as promises and orders. If I am right, however, it is difficult to avoid such attributions without ascribing some sort of promissory obligations and rights. Moreover, users of ChatGPT most likely *want* their exchanges with chatbots to be normatively loaded if only for the sake of creating a common ground with mostly true statements and the ability to hold the chatbot accountable for not contributing a true statement for example (cf., Dennis & Löhr, 2025).

The interpersonal nature of speech acts gives rise to a peculiar form of what I call the *speech act responsibility gap*. In the case of self-driving cars or autonomous weapon systems, it is not clear who can be held accountable for an event in the impersonal sense of the term accountability. In the case of artificial speech, it is not clear who can be held accountable for a false assertion or a broken promise. We are looking for someone to be held accountable but the appropriate target (the chatbot) does not seem capable of it and nobody else seems to be accountable. Moreover, in the case of speech acts, we don’t just expect to hold someone accountable. It seems that the nature of speech acts metaphysically depends on such accountability ascriptions. This is because it is difficult to even make sense of the practice of promising or asserting without ascribing normative relations to an agent who makes or accepts the speech act.

But how can a robot have such directed obligations toward the user to keep a promise, fulfill the order, or make true statements? It seems even more difficult to make sense of directed duties in the case of chatbots than in other instances where AI caused harm. What and who takes responsibility is far from clear. If the gap cannot be filled, this seems to entitle us to doubt whether AI can make any speech act at all, i.e., it calls into question how we describe and conceptualize chatbots. This gives rise to a *conceptual disruption* (Löhr, 2023). It would mean that chatbots are nothing but *stochastic parrots* (Bender, et al., 2021; Koch, forthcoming). Such machine parrots do not technically speak at all (see Williams & Bayne, 2024 for the notion of proto-assertions) but are used by companies to deceive or mislead us into thinking they speak.

**4 How not to bridge speech act responsibility gaps**

Philip Nickel (2013) and more recently Green & Michel (2022), van Woudenberg et al. (2024), and Butlin & Viebahn (forthcoming) have made important contributions to this topic using the notion of a proxy speaker(see alsoCappelen & Dever, 2021, pp. 132). Although not explicitly framed in these terms,[[3]](#footnote-3) proxy speakers, in this literature, are supposed to solve the problem of speech act responsibility gaps. The idea is that since attributing responsibility to chatbots is challenging or impossible, we should consider chatbots as substitutes or proxies for individuals who *are* capable of making linguistic commitments – human agents. Not the chatbot itself, but the human responsible for the chatbot, makes an assertion or promise *via* the chatbot.

The authors all agree that a proxy agent is necessary to *bridge* (rather than fill) the responsibility gap and to make sense of AI speech. Also interesting for the present paper is that the authors just mentioned take for granted that we have no other way but to interpret the output or behavior of chatbots as making assertions, promises, requests, etc., i.e., as making speech acts – if only speech acts on behalf of a human speaker (or “make-belief” in the case of Mallory, 2023).

For example, Nickel (2013, p. 500) argues:

Ultimate responsibility for artificial speech does not lie with machines, but either with persons or companies, or with nobody at all. It is *absurd* [my emphasis] to hold a computer system accountable for a promise or assertion that it makes.

Similarly, Green & Michel (2022, p. 335) argue:

NLG systems produce proxy speech acts on behalf of the engineers who designed them. (…) these entities (natural and legal persons, groups, and corporations) are thus the ultimate bearers of responsibility for any speech acts that the NLG systems perform.

Van Woudenberg, Ranalli and Bracker (2024, p. 18) argue:

However, even if truth had been a design-principle, then ChatGPT still wouldn’t be capable of asserting. This is because ChatGPT would still be incapable of taking responsibility for what it says. For any norm of assertion, even the ones which require a state weaker than knowledge (e.g., having justification), it’s still the case that to “make an assertion is to confer a responsibility (on oneself) for the truth of its content; to satisfy the rule of assertion.”

Butlin & Viebahn, (forthcoming, p. 14) claim:

In any case, neither LaMDA [the name of a large language model developed by Google] nor Sparrow [a chatbot developed by DeepMind] are good candidates for producing assertions, if sanctions for asserting require (i) that interlocutors keep track of asserters’ credibility, (ii) that asserters are sensitive to this, and (iii) that losing credibility is bad for asserters. This is because element (iii) is not satisfied. Even if humans negatively assess their outputs, deduct credibility points and provide according feedback to the systems, there is no sense in which this is bad for the systems. After all, such systems do not have interests of their own, and as a result nothing can be good or bad for them.

I agree that the proxy model is the most plausible explanation of our interaction with current user interfaces like Apple’s Siri or Amazon’s Alexa. When buying something from Amazon via Alexa, it is not that we owe money to Alexa or made a contract with the chatbot. Instead, we engage in a normative relationship with Amazon using Alexa as an interface.

However, none of these interactions necessarily involve speech acts or any actions at all. We can easily make sense of them in the way we make sense of selecting what we want from a website and clicking “order”. Where I disagree with the majority view in the philosophy of linguistic AI, is that this model is not how we should make sense of conversations with more sophisticated large language models. What I disagree with is the generalization of the following overly simplistic model of speech acts of some sophisticated LLMs: Real human beings give their permission to allow the chatbot to make a contract for example on their behalf. According to this model, we only make and accept promises via the chatbot as a proxy. It is the human agent we hold accountable if something goes wrong. It is the human agent who, strictly speaking, makes the speech act.

I argue that the proxy agent account is not sufficient to explain the phenomenon of AI speech. It is also not sufficient to bridge the responsibility gap that is specific to chatbots involving not just impersonal moral or legal responsibility but second personal relational accountability(Löhr, forthcoming). Something is missing when thinking about our interactions at least with the most sophisticated current and probably most future chatbots or virtual assistants (like Alexa 2.0 or even new versions of ChatGPT and Replika), which interact with us all the time (in the case of a voice-based user interface) and that adapts to our use and personality (AI that still learns).

Once we are interacting with LLMs that are still learning *after* the user acquired them, there will inevitably be speech act responsibility gaps that can neither be attributed to the company nor the engineers who build the chatbot. Current scholarship on our normative relationship with chatbots and robots is focused too much on *im*personal legal and moral kinds of responsibility while neglecting much more common and more mundane forms of *inter*personal normativity, which arise in the case of speech.

The key piece of evidence against the proxy account is the following observation: When interacting with robots communicatively, the entitlements or obligations of the person who is legally responsible for the robot’s output (say, Amazon or the private owner) and the necessary duties and entitlements ascribed to the system making the utterance (the chatbot) can diverge.

Again, the company is certainly responsible for *some things* of what Alexa is *saying* (we can compare this to the normative relation between a parent and a child, cf., Tigard, 2021). They will take responsibility if the chatbot makes sexist or racist speech acts.[[4]](#footnote-4) The company will apologize for their chatbot and probably update the algorithm. This is because they *are* responsible and are also in control of fixing the problem. If you tell Alexa to order chocolate, you will order the chocolate from Amazon and not from Alexa. You did not enter a legal relationship with Alexa. Alexa does not have any real moral or legal responsibility to send you chocolate. Amazon does. If establishing a contract is making and accepting a promise (e.g., Fried, 2015, but see Gilbert, 2018), then Amazon promises you the chocolate and you promise to pay for it in return. Like a lawyer, Alexa will mediate between you and Amazon, but not ultimately be the one you establish a contract with. Maybe something goes wrong with the order and instead of chocolate, you get socks. Then it is not that Alexa owes you an apology. That *would* be “absurd” to use Phillip Nickel’s expression. It is Amazon that does. You complain to Alexa, but Alexa is merely a proxy agent for Amazon.

However, for the most part, Alexa will make utterances that Amazon is certainly *not* responsible for. Most of the utterances Alexa makes at least sound like they are constatives, commissive, or even directives, which we cannot reasonably hold Amazon responsible for. Amazon is responsible for its chatbot but not for many of the specific speech acts it makes.

For example, if Alexa tells you that the Sun is shining and this turns out to be a mistake, it is not appropriate to hold Amazon responsible. The reason for this takes us back to the original reason for responsibility gaps pointed out by Matthias (2004): Part of the AI’s functionality is that it learns (recall Matthias’ example of a learning elevator). Amazon gave you a learning robot and they will be right to tell you that you need to give Alexa time to adjust. As Matthias pointed out, the engineer who built Alexa, however, has no control over what it is uttering. Thus, it makes no sense to have reactive attitudes like resentment and indignation (Strawson, 1962/2008; Tollon, 2023) toward the engineer who built the chatbot or the company who sold it to you.

Of course, Amazon becomes responsible for the behavior of Alexa when bad behavior keeps happening after months when Alexa was supposed to have learned to provide the correct information. However, then Amazon becomes liable for a defective product, which is essentially a broken promise that the product will not consistently provide false information. Unless the product is defective (e.g., makes racist utterances) in a way that breaches Amazon’s contractual obligations toward you, it seems that your reactive attitudes regarding the utterance of the chatbot are not appropriately directed at the company.

To illustrate the argument further, imagine that you tell Alexa to turn the oven on at exactly 6 pm. When you come home, the food in the oven is cold and you point this out to Alexa. Again, filing a complaint against Amazon or requesting a refund seems futile. It appears that Amazon is not responsible for your cold food, nor do they have the authority to resolve the issue. They would be responsible for having sold you a defective device – but the device is not defective. It still needs to learn! If Amazon technicians visit and discover no problem, they’ll reiterate that Alexa occasionally makes errors and must learn from them (just like Matthias mentioned for learning elevators in 2004). So, there are errors or things that did not work out, but it is clear who is *not* responsible for it: Amazon is not to be held responsible. The technician also can’t force the machine to learn what you need it to learn. It needs to learn this on its own.

Now imagine that there is a new Amazon product you want to buy, and Alexa “promises” to buy it for you once it becomes available. Alexa fails to do so and now you want to have a reason from Amazon. Amazon lacks access to the reasons behind Alexa’s failure to order the new product. Instead, it can only determine whether the underlying program is functioning correctly. Alexa serves as a black box, concealing the relevant components. It is a learning algorithm Amazon lacks transparency into the precise process behind its decisions. This could be due to an error in the programming, prompting a necessary fix. However, if they fail to identify any issues with the program, it becomes a matter of Alexa’s continuous learning. If they lack control over the problem and the machine functions optimally, we encounter a gap that humans cannot fill or bridge.

**5 What is special about speech act responsibility gaps?**

I hope that at this point, the reader is convinced that speech act responsibility gaps are real and cannot be bridged by humans or corporations. Moreover, it seems that speech act responsibility gaps are different from previously discussed responsibility gaps. One key difference between this *speech act responsibility gap* and the traditional examples is that we have to do with cases that do not necessarily involve physical damage or harm. It seems that the wrong that chatbots can do is relatively harmless in the sense that it mainly applies to harm that can be inflicted by language.

The main difference is that the notion of responsibility that has been discussed so far in the literature on responsibility gaps is mostly *impersonal* (strangers, too, can hold me accountable). When I cause an accident with my car due to being negligent or commit a war crime, then I don’t only wrong the person I harm and their family. I also can be held accountable by the entire community. Driving under the influence of alcohol for example can be a matter of both tort law and criminal law. I am accountable and answerable not just to the victims but also to society at large even if most members of said society are not directly affected by my action. This kind of impersonal responsibility is what is arguably missing when self-driving cars cause an accident or autonomous weapons a war atrocity.

Most of the normative relations discussed so far in the case of speech acts, however, are inherently second-personal or directed at individual speakers. They always involve specific individuals who interact with one another and hold each other accountable. When making a promise I owe the promised action to you but not to the rest of the community. When I make an assertion to you, you could hold me accountable if my assertion turned out to be false and I have to defend myself *to you* but not to the rest of the community (normally at least).

But how can we understand this normative relationship if one of the “agents” is not an agent at all? If linguistic interactions are inherently or at least almost always normatively charged, how can we ascribe entitlements and commitments to artificial systems? This, I argued, creates a speech act responsibility gap that clearly meets the conditions of a standard responsibility gap. As mentioned, according to Köhler, et al. (2018, p. 54), responsibility gaps occur in situations where we would normally look for someone to criticize or sanction (where such actions play a clear societal function) but where no candidate could appropriately be held accountable. And this is precisely what can be observed in conversations with chatbots. We expect to hold someone accountable but there is no agent who we could hold responsible.

Finally, this responsibility gap can be observed in a very strong form, i.e., in a way that avoids many of the worries raised against the very existence of a responsibility gap by Tigard (2021) and Köhler et al. (2017). These authors objected that most of the paradigmatic examples like the automatic elevator or the UBER case are better construed as mere accidents, events where we should not look for anyone to be held responsible. This is not possible in our case. To be able to understand and describe what happens in a conversation with a chatbot, we need to use the only language we have – the chatbot makes statements, promises, demands, criticism, and excuses. However, these speech acts entail normative commitments and entitlements except that there is nobody to whom we could appropriately ascribe these normative relations. No human agent can be held responsible for these speech acts.

To emphasize the last point, I take the speech act responsibility gap to highlight a dilemma that is difficult to ignore: Either we need to extend the notion of responsibility to certain chatbots, or most artificial speech is completely mysterious. This problem goes beyond regular responsibility gaps as it is not just about finding someone to blame. It is about understanding what conversations with chatbots are in the first place. We cannot help but describe Alexa as having ordered food or having said that the sun is shining but if the skeptics like Nickel or Green and Michel are right, unless these utterances can be ultimately ascribed to a human proxy agent, then Alexa has not said or ordered, anything at all.

The speech act responsibility gap is generated by the very same factors that were pointed out in Matthias’s original treatment of the topic. No human speaker is in control over the chatbot’s specific output, it cannot be predicted and, most importantly, the product is still a learning product that changes its nature once unpacked and engaged with. Again, since it is part of its proper functionality that it needs to make mistakes, it needs to learn, not the engineer, the company or any human can be held accountable for mistakes the chatbot makes. If nobody at all can be held accountable then it seems that the chatbot creates no speech at all, which also means that we have no language to describe what it actually does. In other words, we completely lack the language to describe what the chatbot does.

**Conclusion**

I pointed out a new kind of responsibility gap that arises from our interactions with such sophisticated chatbots. The main differences between this *speech act responsibility gap* and the traditional examples are a) that we have to do with cases that do not necessarily involve physical damage or harm and b) that we have to do with interpersonal and directed normative relations rather than a more general impersonal kind of responsibility. The latter is most relevant when, say, a self-driving car injures a pedestrian, or an autonomous gun “commits” a war crime. In the case of promises and assertions, the relevant normative relation is interpersonal. We owe it to the other speakers to keep our promises and make true assertions and only they are entitled to criticize us for our speech acts (exceptions aside).

I further argued that the speech act responsibility gap cannot be bridged by a proxy agent account. Such accounts might work for impersonal kinds of responsibility gaps but not the ones of the communicative kind we encounter for chatbots. The speech act responsibility gap arises, I argued, from the fact that it is difficult to ascribe accountability to the chatbot. This generates a dilemma: We can either deny that the chatbot is accountable, which would diminish its functionality, or accept that we have no idea, no language to describe what the chatbot *is* doing. The speech act responsibility gap is arguably deeper than traditional impersonal responsibility gaps.

**References**

Austin, J.L., (1962). *How to do Things with Words*, 2nd edition, J.O. Urmson and M. Sbisá (eds.), Cambridge, MA: Harvard University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Boge, F. J. (2022). "Why Machine Learning Is Unfair and Black Boxes Are Hard to Open." European Journal for Philosophy of Science, 12(2), 1-25.

Buckner, C. (2019). "Deep learning: A philosophical introduction." Philosophy Compass, 14(10), e12625.

Butlin P. & Viebahn, E. (forthcoming). AI Assertion. *Ergo*.

Brandom, R. (1983). Asserting. *Noûs*, 637-650.

Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering*. Oxford University Press.

Cappelen, H., & Dever, J. (2021). *Making AI intelligible: Philosophical foundations*. Oxford University Press.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, 26(4), 2051-2068.

Danaher, J., & Nyholm, S. (2021). Automation, work and the achievement gap. *AI and Ethics*, 1(3), 227-237.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, 18(4), 299-309.

Darwall, S. (2013) “Bipolar obligation.” In: *Morality, authority, and law: Essays in second-personal ethics I*, ed. S. Darwall, pp. 20–39. Oxford University Press.

Dennis, M. J. (2022). Social robots and digital well-being: how to design future artificial agents. *Mind & Society*, 21(1), 37-50.

Fried, C. (2015). *Contract as promise: A theory of contractual obligation*. Oxford University Press, USA.

Geurts, B. (2019). “Communication as Commitment Sharing: Speech Acts, Implicatures, Common Ground.” Theoretical Linguistics 45 (1-2): 1–30.

Gilbert, M. (2018). *Rights and demands: A foundational inquiry*. Oxford University Press.

Green, M., & Michel, J. G. (2022). What might machines mean?. Minds and Machines, 32(2), 323-338.

Habib, A., (2022). Promises. *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/win2022/entries/promises/>.

Kasirzadeh, A., & Gabriel, I. (2023). In conversation with Artificial Intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2), 1-24.

Koch, S., Löhr, G., & Pinder, M. (2023). Recent work in the theory of conceptual engineering. *Analysis*, *83*(3), 589-603.

Koch, S. (forthcoming). Babbling stochastic parrots? A Kripkean argument for reference in large language models. *Philosophy of AI*. 10.18716/zf3d7793

Koops, B. J., Hildebrandt, M., & Jaquet-Chiffelle, D. O. (2010). Bridging the accountability gap: Rights for new entities in the information society. *Minn. JL Sci. & Tech*., 11, 497.

Köhler, S., Roughley, N., Sauer, H. (2017) Technologically blurred accountability? In: Ulbert, C., et al. (eds.) *Moral agency and the politics of responsibility*. Routledge.

Lewis, D. (1979). *Scorekeeping in a language game* (pp. 172-187). Springer Berlin Heidelberg.

List, C. (2021). Group agency and artificial intelligence. *Philosophy & technology*, 34(4), 1213-1242.

Löhr, G. (2022). Linguistic Interventions and the Ethics of Conceptual Disruption. *Ethical Theory and Moral Practice*, 1-15.

Löhr, G. (2023). Conceptual disruption and 21st century technologies: A framework. *Technology in Society*, 74: 102327.

Löhr, G. (2024). If conceptual engineering is a new method in the ethics of AI, what method is it exactly?. *AI and Ethics*, 4(2), 575-585.

Löhr G. (forthcoming). Why Proxy Accounts of AI Speech Acts Fail. *American Philosophical Quarterly.*

Löhr, G., Dennis, M. Prudential reasons for designing entitled chatbots: How robot "rights" can improve human well-being. AI Ethics (2025). https://doi.org/10.1007/s43681-025-00676-x

Mallory, F. (2023). “Fictionalism about Chatbots.” Ergo: An Open Access Journal of Philosophy 10 (38). <https://doi.org/10.3998/ergo.4668>.

Marsili, N., & Green, M. (2021). Assertion: A (partly) social speech act. *Journal of Pragmatics*, 181, 17-28.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, 6, 175-183.

May, S. C. (2015). Directed duties. *Philosophy Compass*, 10(8), 523-532.

Nickel, P. J. (2013). Artificial speech and its authors. Minds and Machines, 23(4), 489-502.

Nyholm, S. (2018a). The ethics of crashes with self-driving cars: A roadmap, II. *Philosophy Compass*, 13(7), e12507.

Nyholm, S. (2018b). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. Science and engineering ethics, 24(4), 1201-1219.

Nyholm, S. (2023). *This is Technology Ethics: An Introduction*. John Wiley & Sons.

Oimann, A. K. (2023). The Responsibility Gap and LAWS: a Critical Mapping of the Debate. *Philosophy & Technology*, 36(1), 1-22.

Krueger, J. & Osler, L. (2022). Communing with the dead online: chatbots, grief, and continuing bonds. *Journal of Consciousness Studies*, 29 (9-10), 222-252

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, 34, 1057-1084.

Searle, J. R (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.

Searle, J. (2010). *Making the social world: The structure of human civilization*. Oxford University Press.

Shapiro, L. (2020). Commitment accounts of assertion. *The Oxford handbook of assertion*, 73-97.

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy*, 24(1), 62-77.

Strawson, P. F. (2008). *Freedom and resentment and other essays*. Routledge.

Stalnaker, R. (2002). Common ground. *Linguistics and philosophy*, *25*(5/6), 701-721.

Sullivan E. (2022). Inductive Risk, Understanding, and Opaque Machine Learning Models. Philosophy of Science. 89(5):1065-1074. doi:10.1017/psa.2022.62

Swoboda, T. (2018). Autonomous Weapon Systems-An Alleged Responsibility Gap. In V. Müller (Edt.). *Philosophy and Theory of Artificial Intelligence* (pp. 302-313). Springer International Publishing.

Sterken, R. K. (2020). Linguistic interventions and transformative communicative disruption. In: Herman Cappelen, David Plunkett & Alexis Burgess (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press. pp. 417-434

Thomasson, A. L. (2017). Why we should still take it easy. *Mind*, 126(503), 769-779.

Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology*, 34(3), 589-607.

Tollon, F. (2023). Responsibility gaps and the reactive attitudes. *AI and Ethics*, *3*(1), 295-302.

van Woudenberg, R., Ranalli, C., & Bracker, D. (2024). Authorship and ChatGPT: a Conservative View. Philosophy & Technology, 37(1), 34.

Veluwenkamp, H., & Hindriks, F. (2024). Artificial agents: responsibility & control gaps. *Inquiry*, 1-25. https://doi.org/10.1080/0020174X.2024.2410995

Wallace, R. J. (2019). *The moral nexus* (Vol. 8). Princeton University Press.

Williams, I. and Bayne, T. (2024). “Chatting with Bots: AI, Speech Acts, and the Edge of Assertion.” Inquiry: 1–24. <https://doi.org/10.1080/0020174X.2024.2434874>.

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2), 265-288.

1. Others have added interrelated concepts like achievement (Danaher & Nyholm, 2021), accountability (Koops et al., 2010) and retribution (Nyholm, 2018; Danaher, 2016) to the list of disrupted concepts, i.e., concepts that are undermined or challenged by new technologies in fundamental ways (cf., Löhr, 2022). [↑](#footnote-ref-1)
2. I am not claiming to have summarized the large literature on assertion and promises here. I am not claiming that we can’t possibly make sense of speech acts in non-normative terms. All I am saying is that they normally have normative consequences even if these consequences are not necessary for an utterance to count as a promise for example. [↑](#footnote-ref-2)
3. Given that many of the authors have a background in philosophy of language rather than the philosophy of technology, with Nickel as an exception. [↑](#footnote-ref-3)
4. Famously, this happened to Microsoft’s Tay: <https://en.wikipedia.org/wiki/Tay_(chatbot)> [↑](#footnote-ref-4)