Minds and Machines (2022) 32:323–338
https://doi.org/10.1007/s11023-022-09589-8

ORIGINAL ARTICLE

What Might Machines Mean?

Mitchell Green[1] · **Jan G. Michel**[2]

Received: 1 January 2021 / Accepted: 16 January 2022 / Published online: 27 January 2022

Abstract

This essay addresses the question whether artificial speakers can perform speech acts in the technical sense of that term common in the philosophy of language. We here argue that under certain conditions artificial speakers can perform speech acts so understood. After (§1) explaining some of the issues at stake in these questions, we (§2) elucidate a relatively uncontroversial way in which machines can communicate, namely through what we call verbal signaling. But verbal signaling is not sufficient for the performance of a speech act. To explain the difference, we (§3) elucidate the notion of a speech act developed by Austin (*How to Do Things with Words*, 1962) in the mid-twentieth century and then discuss Strawson's ("Intention and Convention in Speech Acts", 1964) influential proposal for how that notion may be related to Grice's ("Meaning", 1957) conception of speaker meaning. We then refine Strawson's synthesis in light of Armstrong's ("Meaning and Communication", 1971) reconceptualization of speaker meaning in terms of objectives rather than intentions. We next (§4) extend this conception of speech acts to the cases of recorded, proxy, and conditional speech acts. On this basis, we propose (§5) that a characteristic role for artificial speakers is as proxies in the performance of speech acts on behalf of their human creators. We (§6) also consider two objections to our position, and compare our approach with others: while other authors appeal to notions such as "quasi-assertion," we offer a sharp characterization of what artificial speakers can do that does not impute intentions or similarly controversial powers to them. We conclude (§7) by raising doubts that our strategy can be applied to speech acts generally.

Keywords Artificial speaker · Proxy speech acts · Signaling · Illocutionary act · Assertion · Commitment

✉ Jan G. Michel
jan.michel@hhu.de

Mitchell Green
mitchell.green@uconn.edu

[1] University of Connecticut, Storrs, USA

[2] University of Düsseldorf, Düsseldorf, Germany

1 Introduction

As intelligent machines occupy an increasingly pervasive role in human life, they take on an ever-expanding set of functions previously reserved for human beings and domesticated animals. Among the most recent achievements of intelligent machines are forms of communication that bear striking affinity to what philosophers of language call *speech acts*: acts such as asserting, warning, promising, and requesting, which exhibit the 'saying makes it so' property, that is, the property of being an act that may be performed by saying and meaning that one is doing so (cf. Green, 2020, 2021a).

The question whether machines can perform speech acts matters for at least two reasons. First, speech acts are widely viewed as the building blocks of human conversations. Different approaches to conversation (Grice, 1975; Roberts, 2018; Stalnaker, 2014, etc.) converge in seeing conversations as built out of sequences of speech acts. Accordingly, if machines can perform speech acts, that will move them that closer to being able to converse with human beings in a full sense.

Second, it is commonplace to assess machines in normative terms. This we do whenever we ask if a machine is performing as it should, that is, as it is designed to perform. However, if machines can perform speech acts, that fact may open a new dimension of normative assessment, namely moral evaluation.[1] The reason is that some speech acts put their producer in a moral relationship to their addressee: those who, for instance, make promises they have no plan of fulfilling are subject to censure. Conversely, we would expect a machine capable of speech acts also to be able to engage with them as a recipient. In order to be such a recipient, the machine in question needs to be capable in principle of uptake, that is, it must at a minimum be able to recognize that a speech act of a certain kind—such as a promise—is being performed. The capacity for uptake would in turn enable us to undertake commitments to some machines, and doing so could in turn put us under moral obligations to them.

The ability of machines to participate as producers or recipients of speech acts thus opens up the possibility of weaving them into our moral fabric. That, however, is in tension with a widely shared sense that machines do not have moral standing, or at least not in the way that human persons do. In order to help resolve this tension, we proceed as follows. We first (§2) elucidate a relatively uncontroversial way in which machines can communicate, namely through what we call verbal signaling. But verbal signaling is not sufficient for the performance of a speech act. To explain the difference, we (§3) elucidate the notion of a speech act developed by Austin (1962) in the mid-twentieth century and then discuss Strawson's (1964) influential proposal for how that notion may be related to Grice's (1957) conception of speaker meaning. We then refine Strawson's synthesis in light of Armstrong's (1971) reconceptualization of speaker meaning in terms of objectives rather than intentions. We

[1] Ethical normativity is one among a larger class of types of normativity which also includes norms of etiquette, practical rationality, and theoretical rationality among others. Therefore, not every normative assessment is also a moral evaluation.

next (§4) extend this conception of speech acts to the cases of recorded, proxy, and conditional speech acts. On this basis, we propose (§5) that a characteristic role for artificial speakers is as proxies in the performance of speech acts on behalf of their human creators, and that some artificial speakers meet the cognitive and conative conditions for performing speech acts in this role. We (§6) also consider two objections to our position, and compare our approach with others: while some authors appeal to notions such as "quasi-assertion," we offer a sharp characterization of what artificial speakers can do that does not impute intentions or similarly controversial powers to them. We conclude (§7) by raising doubts that our strategy can be applied to speech acts generally.

2 Signaling and Communication in Machines

The simplest machines have functions: levers, pulleys, wheels, and screws do their jobs without engaging in behaviors in pursuit of a goal. More complex machines do just this: heat-seeking missiles and Roombas pursue goals because they modify their behavior in response to environmental changes, and continue so to modify it until a goal is reached. To best explain the behavior of these machines, however, we need not assume that they have such mental representations as objectives, desires, or plans; it is sufficient to attribute goals to them.[2]

A great variety of machines also communicate. (Odometers, EKG machines, tire-pressure gauges, satellites, and doorbells are cases of this kind.) To account for their communicative behavior, we may apply a conceptual framework provided by the theory of signaling as found in literature on the evolutionary biology of communication. According to that literature,[3] some behaviors and traits are *signals* in the sense of being designed to convey information. Although one way to design something is to do so intentionally, the notion of design need not be construed solely in terms of intentions. The bright coloration on the skin of an Amazonian tree frog signals (i.e., designedly conveys the information) that the frog is toxic. My scowling face signals that I am angry even if I do not produce that scowl intentionally. Human artifacts such as colored bricks on a sidewalk likewise perform a signaling function, such as that a certain area is for cyclists only. When S is a signal, we may always ask what it signals, and we may ask how it signals what it does. Also, S may be a signal for one type of audience but not for another. Bright anuran coloration may be a signal of toxicity for snakes and birds, but not for other species that are immune to their poison or do not eat frogs. Further, 'information' as used here is not factive, and it is possible to signal what is not the case. This would occur when, for instance, a mutant tree frog is born into a population of frogs whose bright coloration signals their toxicity:

[2] Ascribing a goal to a system does not require a mental ascription. We may for instance ascribe goals to plants (to capture more sunlight, for instance) without ascribing mental representations to them.

[3] A *locus classicus* is Maynard-Smith and Harper 2004. See Green 2019 for an elucidation of some of the philosophical significance of this literature.

this mutant, however, is brightly colored but not toxic. About this mutant we might say, "Its bright coloration means that it is toxic, but it isn't."

Machines signal by virtue of being artifacts designed to convey information, as opposed to being results of natural selection. They communicate when such signals are received and correctly interpreted. NLG (Natural Language Generation) machines have the further capacity to signal with language.[4] Linguistic items such as phrases are not themselves signals, but they are designed to be so used. Further, linguistic items with semantic content are designed to signal in a way that is constrained by that content. Thus an indicative sentence S whose truth conditions are ⟦S⟧ is designed to be used as a signal that the actual situation is a member of ⟦S⟧.[5] We may put this succinctly via the idea that indicative sentences encode information, and that they do so in a manner constrained by their semantic content. Building on the work of Green (2021b), we may accordingly extend the notion of a signal to verbal cases along the following lines:

Sufficient Condition for Verbal signaling (SCV):

> An NLG system that is designed to convey the information encoded by the sentences it tokens, and which tokens indicative sentence S, thereby verbally signals the information encoded by S.

In parallel with what we noted about the relation of signaling and communicating, we should expect verbal communication to occur when verbal signaling is successful. Such forms of signaling and communicating, however, are a far cry from performing speech acts such as promises and assertions. One reason is that such speech acts are widely understood as requiring intentions, which many of us hesitate to ascribe to even the most sophisticated machines. It may seem, then, that even if they verbally signal and communicate, machines do so in a dramatically different way from what we find among human beings.

3 Minds, Language, and Machines

To assess this apparent gap between machine and human communication, let us do some ground-clearing in the philosophy of mind. Mental states may be distinguished into four broad types: cognitive, conative, affective, and experiential (Green, 2018). *Cognitive* states include beliefs, memories, and expectations, and represent the world as being a certain way.[6] *Conative* states include intentions, desires, objectives, and plans, and represent the world as to be made a certain way.[7] Conative states tend, in conjunction with cognitive states, to produce actions that can be reconstructed

[4] Natural Language Generation (NLG) systems are capable of generating understandable texts in human languages, typically starting from some non-linguistic representation of information as input, cf. Reiter & Dale 2000.

[5] For current purposes it suffices to construe '⟦S⟧' as referring to a set of possible worlds. See Green 2021a for further discussion.

[6] Some authors characterize this in terms of the notion of *mind-to-world direction of fit*. See for instance Searle 1983.

[7] Searle (1983) characterizes this in terms of a *world-to-mind direction of fit*.

as rational. ("A grabbed an umbrella as she left home, because she believed it was raining outside, wanted to stay dry, and believed that the umbrella would help her do that.") *Affective* states include emotions and moods, where the former differ from the latter in mandating an intentional object. *Experiential* states are those possessing a phenomenal character (Michel, 2011), and are exemplified in sensation, imagery, and some types of introspection. While some affective state-types have a phenomenal character (there are distinctive ways that rage and grief feel), neither cognitive nor conative states necessarily have phenomenal character: there need not be anything it is like to believe that $2+2=4$ or to plan to learn Japanese before the age of 50. Furthermore, we see no conceptual, metaphysical, or nomological necessity for the four foregoing types of mental state to co-occur: there could be entities with cognitive and conative states but no affective or experiential states, for instance. While lacking the capacity for happiness, love, or fear, such entities may still navigate the world effectively. They could also perceive it, although the process would be akin to blindsight if they possess vision, "deafaudition" if they hear, etc.

Against this backdrop, we turn to the philosophy of language. Austin begins *How to Do Things with Words* (Austin, 1962) with a distinction between constative and performative utterances. The former are descriptions of a putatively independent reality, such as 'It is raining outside' or 'Dodoma is the capital of Tanzania.' In performative utterances, by contrast, our words also constitute a change in the world: 'I promise to visit you in Dodoma', if said under the right conditions, creates an obligation on the part of the speaker to visit her addressee in Dodoma. On closer scrutiny, however, Austin finds that the constative/performative distinction needs to be replaced, since even constative utterances (a central example of which is assertion) effect a change in the world. For instance, someone who asserts that Dodoma is the capital of Tanzania undertakes a commitment to the truth of that claim, together with a liability to be accused of error if that claim is false and of malfeasance if it is unjustified or a lie (Green, 2016).

By the end of *How to Do Things with Words*, Austin has replaced the constative/performative distinction with a tripartition among locutionary, illocutionary, and perlocutionary acts.[8] Our main interest in what follows will be illocutionary acts, but for clarity we hasten to distinguish them from locutionary acts, which are acts of uttering sentences with a definite semantic content. Illocutionary acts (now often referred to as *speech acts*) are acts of a sort that can, but need not be, performed by saying and speaker-meaning that one is doing so (Green, 2020; the notion of speaker meaning is explained in the next paragraph). Promising is an illocutionary act on this criterion, but so is asserting, as one can assert P by saying and speaker-meaning that one is doing so.[9] Note that one can perform a locutionary act without performing an illocutionary act: the somniloquist who utters, "I promise to visit you …," does not make a promise, and the actor who says on stage, "Alas, poor Yorick! I

[8] Sbisà (2007) lucidly documents Austin's change of strategy over the course of his lectures.

[9] One can, of course, also assert without saying that one is asserting. We should mention that on the present criterion for being a speech act, convincing and offending are not illocutionary acts (since they do not pass the 'saying makes it so' test). Austin would term them *perlocutionary acts*.

knew him, Horatio," does not assert that he knew anyone named Yorick even as he portrays a character who does exactly that.

At about the same time that Austin was developing his theory of speech acts, Grice (1957) was articulating a theory of non-natural meaning (now known as *speaker meaning*) as depending on complex sets of intentions. For Grice, to mean (sensu speaker meaning) that p, one needs to make an utterance with an intention to produce a psychological effect on an addressee, with the further intention that that effect be produced in part through the addressee's recognition of one's intention. (This complex intention is known as a *reflexive-communicative intention*.)

Grice's work was widely recognized as a breakthrough, but it raised the issue of how speech acts and speaker meaning relate to one another. Strawson (1964) defends an answer to this question by, first, pointing out that Austin's claim (Austin, 1962, p. 14) that all speech acts depend on extra-semantic conventions is overblown: some speech acts are convention-dependent in this way, but many are not.[10] For those speech acts not depending on extra-semantic conventions, Strawson proposes that they be understood in terms of speaker meaning. One asserts that P, for instance, by uttering a sentence whose conventional meaning is P, and with the further intention of producing a belief in an addressee by means of their recognition of one's intention.[11] No extra-semantic conventions need be involved.

Armstrong (1971) further refines Strawson's synthesis of Austin and Grice by pointing out that intentions are needlessly demanding for the role they play in speech act theory. Instead, Amstrong proposes replacing the notion of intention with that of an *objective*. His reason is that unlike an intention, one can have an objective of doing A with little confidence of being able to do it. Armstrong also observes that we may recast Grice's conception of speaker meaning in terms of objectives with no loss of insight into speech acts.

Replacing intentions with objectives in an account of speech acts is salutary for our present project as well. The reason is that some types of intention appear to be bound up with consciousness. In particular what some authors call *pure intention* (Setiya, 2018), which is intending not accompanied by any action, seems to be a mixture of mental imagery, inner speech, and impulses, all of which are typically consciously experienced. By contrast, 'objective' does not carry that imputation. This opens up the possibility of entities lacking consciousness acting with the objectives required to perform speech acts.

To justify the ascription of objectives and not merely goals to machines, however, we do well to search for cases that harbor mental representations. Developments in AI over the last half-century have made a powerful case for the attribution of mental representations to machines. When such representations (a) take propositional form ("The marble is in the bowl"), (b) occur in a system that treats them

[10] In pronouncing a couple married, a priest invokes conventions that transcend those giving our words their conventional meanings: such conventions are thus extra-semantic in our sense. Also, some illocutionary conventions appear to have a "strict liability" character and thus do not require much in the way of speaker intentions. Thus in the Sunni Muslim practice of "triple talaq," a husband who utters 'talaq' three times in front of his wife, thereby divorces her (Ahmad 2009) regardless of what he intends.

[11] Bach and Harnish 1979 develop this approach in further detail.

as corresponding to how the world is, (c) that system contains sufficient internal complexity and perceptual competence to justify attribution of concepts that would underwrite discrimination of marbles from, say, dice and cumquats, and (d) that system's internal complexity also underwrites such inferences as 'The marble is in the bowl' to 'The marble is not in the cup'; when these four conditions are met, then we may reasonably describe such representations as beliefs and thus as cognitive states. Similarly, when such representations (a*) take propositional ("The marble should be put in the bowl") or imperatival ("Put the marble in the bowl") form, but (b*) occur in a system that treats them as corresponding to how the world is to be modified, while also meeting analogues of conditions c and d above, then we may describe them as objectives, and thus as conative states.[12]

Thus far, in order to fathom the apparent gap between machine and human communication, we have observed that speech acts may be underwritten by objectives rather than intentions, and that such objectives may be had by systems lacking consciousness. We have also adumbrated the conditions that could justify the attribution of certain mental states to machines. We next argue that some such machines can perform illocutionary acts.

4 Recorded, Proxy, and Conditional Speech Acts

The Armstrong-Strawson synthesis of Austin and Grice provides a basis for further refinements in the theory of speech acts; some of these refinements also help us make progress on the question whether machines can illocute.[13] First of all, and for the easiest case, we note that agents may *record* their speech acts for wider transmission than is normally possible with spoken discourse. Writing as well as voice and video recordings are cases of this kind. Thus when an agent writes, 'Dodoma is the capital of Tanzania,' we take the inscription to record her dateable utterance. Although we do sometimes describe inscriptions as "saying" that soandso, this is loose talk, as there is no question of inscriptions performing speech acts; likewise for voice or video recordings of speech acts.

Second, an agent or group of agents may also employ a *proxy* to illocute on their behalf. As Ludwig (2020, p. 311) observes, one agent may be authorized to speak on behalf of another individual or group (of which she may but need not be a member). Thus, when a speaker A says under appropriate conditions, "The City Council approves the expenditure of funds for a new bus station," A is not herself approving any such expenditure (after all, she may even have voted against the measure and at any rate isn't authorized to approve any expenditures unilaterally). In a *proxy speech*

[12] We here sedulously avoid the question whether machines have minds, or, in the parlance of Damassino and Novelli, whether they have "true intelligence" (2020, p. 463). Perhaps having a mind or true intelligence requires all four of cognitive, conative, affective, and experiential states, and it may require more sophisticated cognition and conation than we will require of machines in what follows. Neither outcome would threaten our argument.

[13] We follow common practice in using 'illocute' throughout this paper as an intransitive verb meaning 'to perform a speech act'.

act, then, one speaker illocutes *on behalf of* an entity distinct from herself, and it will not follow from the fact that a speaker has done so that the speaker has performed any speech act on behalf of herself, or even that she could do so. However, as Nickel points out, a proxy in such a situation is normally expected to possess the cognitive and/or computational sophistication needed to accurately represent the intentions or objectives of the body on whose behalf she is speaking (Nickel, 2013, pp. 500–501); otherwise there will be little point in using a proxy rather than just recording one's speech act for later playback.

Finally, we have *conditional* illocutions. I could make a bet conditional on certain conditions obtaining: If those conditions do obtain, then I either win or lose money depending upon the terms of the bet. A similar structure emerges with promises and assertions. But suppose that a competent human speaker A asserts Q *conditional* on P's obtaining and that subsequently P does obtain. In October 2020, for instance, no one knew who would win the 2020 US Presidential election, but speaker A asserted that if Biden wins, he will tackle the pandemic.[14] From this, it does not follow that A has *asserted* that Biden will tackle the pandemic. Nevertheless, she is, now that Biden has actually won, *committed* to that claim. For now that he has won, if Biden does not tackle the pandemic, this is good reason for concluding that A was mistaken in performing her original conditional speech act.[15]

Proxy speech acts and illocutionary commitments come together in ways that prove useful for those undertaking the commitments in question. A state government might tax all earners at a rate of 6% of their adjusted gross income. If person N earns $50,000, then the state is committed to demanding that N pay $3000 in taxes, and we are only speaking loosely when we say that the government is demanding that N pay that amount. However, it might behoove the State of Missouri to appoint tax collectors who make such demands on its behalf. Such agents serve as proxies of the government and make demands that accord with its illocutionary commitments: they in effect activate those commitments in the form of illocutions to ensure widespread compliance with the tax code. To that end, a taxpayer might receive a letter such as the following from a tax collector:

> Dear ...,
> The Department of Revenue of the State of Missouri has received your payment of $1,234.00 for your 20XX state income taxes. However, our calculations show that you owed $1,432.00 for 20XX. As a result, you have an unpaid balance of $198.00. You may pay this balance at the following secure website: www.missourirevenue.gov. Failure to pay this balance by 30 June 20XY will result in 10% of your current balance being added to your bill. Please see the

[14] This example is inspired by Malakoff 2020.

[15] A's asserting Q conditional on P's obtaining may be realized in the form of utterance of a conditional, 'If P, then Q.' However, we do not need to take a stand on the linguistic form that conditional illocutions take. Also, the notion of commitment is explicated in different ways by different authors. However, a minimal feature that most explications share is that one who is assertorically committed to proposition P is right or wrong on the issue of P depending on whether P is true. See Green 2016 for further discussion.

attached pages for further explanation of our calculation of the amount you owe for 20XX.

Sincerely,
Missouri Department of Revenue

While we would expect the author of this letter to have produced it with a certain objective, it is doubtful that they need to have reflexive-communicative objectives of the sort normally thought to be required for speech acts such as assertion. Instead, so long as the bureaucrat has been authorized to speak on behalf of the State of Missouri, and acts within the remit of illocutionary commitments of that body, the above letter may be properly read as containing records of assertions, a warning, and a request. This is confirmed by the likely experience of the addressee of this letter, who will feel that she is being *told* she underpaid her state taxes, *warned* to pay the unpaid balance soon, etc. Further confirmation of the illocutionary nature of the acts that the above words record is that some of them are lie-prone: we may readily imagine the author of the above letter to have manipulated figures to make it appear that the taxpayer owes unpaid taxes when in fact she does not. In that case the author of the letter has lied on behalf of the Show-Me State.[16]

We conclude from the foregoing that when a proxy executes illocutionary commitments on behalf of their principal,[17] they do not need to support their acts with reflexive-communicative objectives that would otherwise be required for such speech acts. Instead, so long as they act within the remit of their principal's commitments, they may illocute on that principal's behalf. These findings will guide us as we reflect on the ability of some NLG systems to illocute.[18]

5 NLGs as Proxies for Our Commitments

Human bureaucrats are prone to error and other kinds of "malfunction". As a result, it may behoove an organization to program and install intelligent machines that can automate their tasks. A municipality might for instance set up an entirely automated system, Traffic I, connected to traffic cameras deployed at strategic locations, and through which drivers who are detected going over the speed limit will receive a letter such as the following:

[16] See Stainton (2016) for further discussion of the way in which assertions are lie-prone.

[17] Following Ludwig, we use 'principal' here to refer to the person or group that "asserts something through another (the proxy) who speaks on the principal's behalf" (Ludwig, 2020, p. 307).

[18] Readers familiar with Grice (1957) will recall that he offers an example to argue that mere communicative intentions are insufficient for speaker meaning. This is the "handkerchief" example, in which someone secretly places A's handkerchief at a crime scene to get to the police to believe that A is the culprit. The example shows that speaker meaning must be in some sense overt, a feature we find in the next case that Grice considers, namely that in which Herod presents Salome with St. John's severed head. Our hypothesis as to why the Missouri bureaucrat case is one of speaker meaning is that by invoking the institution of the Missouri Department of Taxation and its associated powers, the bureaucrat's act achieves a kind of overtness that the person who places A's handkerchief at the crime scene lacks.

Dear ...,
Our cameras recorded that you were traveling at 100 km/h on Autobahn 20 at 5:47 pm on 17 March, 20XX. The speed limit for this road is 80 km/h, and as a result you have violated the traffic laws of Mecklenburg-Western Pomerania. For this reason you are required to pay a fine of 50 Euros. A picture of your vehicle taken at the above time and location is enclosed with this letter. You may pay the fine at the following secure website: www.uckermarklandkreis.de. Failure to pay the fine by 1 April, 20XX, will result in another 10 Euros added to your bill. This letter has been generated automatically and is valid without signature.

Sincerely,
Municipality of Uckermark

The machine-generated letter does verbally signal that the recipient owes 50 Euros. However, it may be doubted that Traffic I also asserts that the recipient owes this amount. The reason is that Traffic I lacks the degree of autonomy from its designers required to act with objectives such as we described in §4. Instead, Traffic I is akin to an automated cashier in a grocery store which has goals (such as receiving a certain amount of cash for a set of goods being purchased), but not the mental representations required to underwrite objectives.

One basis for attributing mental representations to a system is that such an attribution is an ineliminable part of the (or a) best explanation of that system's behavior. Moreover, if an essential part of a best explanation of a system's behavior is that we attribute to it the conditions a–d as introduced in §3, then we are justified in attributing to the system beliefs and hence cognitive states. Analogously, if we ascribe a*–d* as part of a best explanation of that system's behavior, we are justified in ascribing objectives and hence conative states. To this end, imagine that the Uckermark municipality switches to a new technology, RobotCop. While Traffic I was primarily a letter-generating system equipped with traffic cameras that sent tickets to speeders, RobotCops are mobile police robots whose main task is to prevent accidents and punish lawbreakers on a certain section of highway in the Uckermark region. RobotCops can move autonomously along the hard shoulder, are equipped with traffic cameras and laser guns that allow them to monitor traffic and measure speeds. They can also continuously monitor the current overall traffic situation on the relevant highway sections and register relevant changes in traffic patterns. In this way, the system comprising such robots maintains a round-the-clock overview of the number of vehicles on the relevant highway sections, their speeds, and their distances from each other. The system is also linked to the official weather forecast for the Uckermark region and is thus informed, for example, about storm warnings and other weather events relevant to traffic. Moreover, the RobotCops have access to traffic statistics databases from the last ten years, which they can use to calculate or predict the times and locations of likely accidents and moving violations.

To illustrate its behavior: a RobotCop catches a speeding driver and reports this to its nearby RobotCop "colleague," whose task then is to stop the speeding driver at a suitable location, check his personal details, and inform him of what he is charged

with and what consequences this may have. However, in order to consider the circumstances that may have led to the driver's behavior, the RobotCop can ask questions about those circumstances, much like a human peace officer. Depending on the driver's answers, the RobotCop has some discretion; for example, it may fine the driver 20 Euros or merely issue a warning.

Rather than construe the exercise of discretion as conscious consideration of alternatives, we note that RobotCops are built on a deep-learning architecture enabling them to autonomously assess different cases within a given range—in our case from a warning up to 20 Euros. Prior to its first deployment, fairness in RobotCop's algorithms has also been sought for the purpose of minimizing bias in its policing.[19] RobotCops as described here behave—with respect to the tasks for which they are deployed—with a sophistication comparable to that of human peace officers. (They meet conditions a–d of §3 above, for instance.) We take it, moreover, that folk-psychological belief/desire-based explanations of behavior are legitimate in the human case.[20] These two observations provide all the reason we should hope for to conclude that belief/desire explanation is legitimate for our law enforcement AIs as well. Indeed, for purposes of our argument, such AIs need only harbor objectives rather than desires. The reason is that a law-enforcement AI such as a RobotCop acts as a proxy for the police department that has deployed it, and, so long as its utterances enact the illocutionary commitments of that department, those utterances will also be speech acts performed on that department's behalf.

Put more formally, our argument is as follows:

1. Some machines (a) verbally signal in their role as proxies on behalf of their principal, and (b) such verbal signals are within the scope of their principal's illocutionary commitments. (Supported by case of Traffic I.)
2. When a machine verbally signals in its role as a proxy on behalf of its principal, and such verbal signals are within the scope of its principal's illocutionary commitments, it illocutes by tokening sentences with communicative objectives, and without reflexive-communicative objectives. (Supported by case of RobotCops.)
3. Some machines illocute by tokening sentences with communicative objectives, and without reflexive-communicative objectives.

The above argument is obviously valid, having the form of a *modus ponens*. To ensure its soundness as well, we note that the quantifier 'some machines' in steps 1 and 3 must be construed as ranging over technologically possible entities rather than any that (so far as we are aware) have been built. Put differently, the above argument establishes that there are technologically possible machines that illocute under certain conditions.

[19] There is a large literature on machine learning addressing these and other issues, see e.g. Goodfellow et al. (2016). In addition, topics such as algorithmic bias, artificial discretion, and responsibility attribution are also receiving increasing attention.

[20] Accordingly, we take it that the possibility of purely neurophysiological explanations of such behavior would not undercut folk-psychological explanations for members of our own species. For part of what makes an explanation the, or at least, a, best explanation, is simplicity; and folk-psychological explanations tend to be dramatically simpler than most neurophysiological explanations of human behavior.

6 Responses to Objections and Comparison with Other Approaches

We have argued that machines such as RobotCops can perform speech acts such as assertions, directives, and warnings. Further evidence that RobotCops perform such speech acts is found in the fact that they may be designed to lie under certain circumstances.[21] The Volkswagen Corporation notoriously designed its vehicles to mask their exhaust emissions when those vehicles detected that they were being tested for precisely this characteristic (Ewing, 2017). We do not claim that these vehicles lied, but the strategy that the VW engineers used provides us with inspiration. For we may imagine the RobotCops designed with the feature that if a motorist that they have pulled over challenges their accusation that they were speeding, the RobotCop replies with the remark, "My radar registered that you were traveling 10 km/h over the speed limit for this area," and they say this whether or not that remark is true, and whether or not they believe it to be. (To keep this malfeasance from being immediately exposed, we can also imagine that the RobotCops perform these strategic lies only 25% of the time, controlled by a random generator.)

The argument's conclusion exposes us to the following two objections. First, a skeptic might attempt to cast doubt on our position by raising the possibility that RobotCops' capacity to illocute is due merely to their being authorized to act as proxies, together with their ability to enact conventionalized speech acts such as we saw in the triple-talaq case of note 10 above. If so, then our argument succeeds "on the cheap," since nearly anything can be deputized to perform conventionalized speech acts. However, while it is true that no minimal cognitive conditions must be met by a potential proxy to be appointed by a principal, it is not the case that all speech acts that the RobotCops perform are conventionalized. For instance, *warning* is not a conventionalized speech act, but is clearly one that RobotCops can perform. Similarly for assertion.[22]

A second criticism concerns the possibility of moral assessment of machines. For, a critic might observe, a system can assert only if in doing so it can also lie. Second, lying is immoral. And third, machines cannot be subjects of moral assessment (except derivatively by being conduits to the moral assessment of their builders). These three premises together imply that machines cannot assert.

In reply, we begin with the observation that the moral imperative against lying is a regulative rather than a constitutive norm.[23] Just as we could have traffic without traffic laws, we could have a community in which assertion is practiced without the moral rule that speakers shall not lie. Instead, there need only be a *norm* proscribing lying, and, as observed in note 1 above, that norm need not be moral. An example

[21] We characterize lying (restricted to the case of assertions) as asserting P in a situation in which one does not believe that P. This characterization does not include the further, and controversial condition that the speaker intends to deceive an addressee. See Krstic (2018) for further discussion. Also, Kneer (2021) reports evidence strongly supporting the conclusion that human subjects are prepared to ascribe lies to machines.

[22] See Green (2021a) for an argument that warning does not rely on extra-linguistic conventions, and Green (2016) for such an argument concerning assertion.

[23] The distinction between regulative and constitutive rules can be found in several places; the most prominent is perhaps Searle (1969, pp. 33–42).

of such a non-moral norm governing assertion is Grice's maxim of Quality ("Do not say what you believe to be false"; 1975, p. 46), which regulates well-conducted conversations. This entails in turn that the critic's premise that lying is immoral is either false on account of being too strong, or too weak to yield the wanted conclusion. That premise could be read either as

(a) "all possible lies are immoral," or
(b) "all actual lies are immoral."

On the first reading, we may see the falsity of the premise by noting the regulative nature of the rule. On the second reading, the case of the RobotCops falls outside the premise's scope since these machines are not actual. On either reading, then, the argument making use of the premise that lying is immoral fails: the argument is unsound on the first reading and invalid on the second reading. Consequently the above argument fails to establish that machines cannot assert.[24]

Finally, let us compare our approach with two others brought forward by Nickel (2013) and Freiman and Miller (2020). We owe to Nickel the insight that NLG systems produce proxy speech acts on behalf of the engineers who designed them (pp. 499–501). Nickel also observes that 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. Nickel is however insufficiently clear on the extent to which NLG systems perform speech acts.

Nickel argues that NLG systems may be regarded as so-called *speech actants* which he defines as follows:

"a speech actant is an entity that produces linguistically meaningful messages for which:

1. The content and force of the message is causally due to the entity and conditioned by its generative inferential and linguistic activity;
2. The message is delivered actively (it is *uttered*);
3. The entity is usually sensitive to the evaluation-conditions for the utterance in the contexts of delivery (e.g., relevance);
4. More specifically, if the entity presents something as true, it is (usually) responsive to relevant evidence, to its other logically related representational and behavioral states, and to the truth; and
5. The message could in principle be insincere, in the sense that it deviated intentionally ('by design') from relevant norms of assertion." (Nickel, 2013, p. 493; emphasis in the original)

It is, however, not entirely clear what Nickel's overall thesis is: while he contends in one place that "some of [the existing NLG technologies] count as speech actants to at least a limited degree" (pp. 493–494), he claims soon thereafter that "existing NLG

[24] Kneer (ibid.) also presents experimental evidence that subjects are prepared to assess machines to which they have ascribed lies as having done something immoral.

systems satisfy the conditions for being speech actants to a substantial degree" (p. 495; see also the abstract to his paper).

Either way, holding that NLG systems can produce speech acts to a limited extent, or to a substantial degree, leaves unsettled the question whether they can perform speech acts. For we may doubt what clear sense may be attached to something's being a speech act in part, or to some degree. Promising, asserting, appointing and betting are qualitative rather than quantitative notions. By contrast, our view that some artificial agents can illocute in their capacity as proxies for other entities, addresses the question unequivocally.

Freiman and Miller (2020) also address the question whether machines can perform speech acts, particularly assertions. After an insightful analysis and critique of Bruno Latour's views on the ability of machines to perform speech acts, these authors argue that quasi-assertion is a type of assertion, while acknowledging that it differs from assertion in various ways (Freiman and Miller, 2020, pp. 428–429). Also, what appears to be their main reason for thinking of machines as capable of making assertions is given in the following passage:

> If a verbal announcement on an airport loudspeaker constitutes an assertion when it is made by a human employee, why doesn't the same verbal announcement on the same loudspeaker constitute an assertion when made by a computer? The function of the message, the explanation of why subjects get knowledge from it, and the phenomenology are the same in both cases. That an employee can be insincere and a computer cannot does not constitute a good reason to distinguish the two in this context. (Freiman & Miller, 2020, p. 428)

This argument seems to be aiming to show that the machine utterance over the airport loudspeaker should count as an assertion. If so, it raises the question, what would be the point of introducing a notion of quasi-assertion at all? Further, to the rhetorical question these authors raise at the end of their first sentence, we would reply as follows: while we do not claim that machines cannot possess intentions, we would refrain from taking a stand on this controversial question. By contrast, Freiman and Miller uncritically accept a functionalist conception of speech acts which begs the question against anyone who might doubt that machines can have intentions. More precisely, that view is question-begging unless they provide, as we have done, an account of how a machine can illocute without possessing intentions.

7 Conclusion

We have argued that under certain tightly constrained conditions, machines can illocute. In the course of that argument, we also showed, against a widely shared consensus in the philosophy of language, that it is possible to illocute without reflexive-communicative intentions.[25]

[25] We expect this result to have far-reaching consequences for other areas in which speech acts play an important role, such as the philosophy of scientific discovery and the role that declarative speech acts play in it (Green (2021b); Michel (2019); Michel (2020); Michel (2022)).

Ours are sufficient, rather than necessary conditions for the performance of illocutionary acts, and we thus leave open the possibility of other conditions under which machines may perform such acts. A different route that might be sought for the extension of our argument to other speech-act types concerns those possessing sincerity conditions consisting in affective states. Can a machine apologize for a mistake it has made, or sympathize with a human user's plight? To be sincere, these acts require the speaker to feel regret or remorse (for apology) and sympathy (for sympathizing). Because these affective states also have a phenomenology, any putative case of apology or sympathizing would be insincere so long as machines are incapable of experiential states. Human users who fail to reflect on these limitations might accept machine apologies and expressions of sympathy at face value, while those who do so reflect will refuse to accept them precisely on the ground of their in principle infelicity. Until they are imbued with phenomenal consciousness, then, machines and their designers may be able to fool some of the people some of the time, but not all of the people all of the time.

We are also in a position to address the tension noted in §1 concerning moral obligations to and by machines. For, first, and as argued in §5, not all speech acts are constitutively governed by moral principles: asserting is not, and presumably certain of the other speech acts likely performed by RobotCops, such as warning and refusing, are not as well. Second, should there be conditions under which a machine can perform speech acts constitutively governed by moral principles, we may follow Nickel (2013) in contending that the moral evaluation of such acts is to be traced back to their designers. Alternatively, and in light of the results of Kneer (2021), we may wish to include that machine among those agents whom we assess morally. These two strategies are of course compatible with one another.

Acknowledgements Research for this article was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, Project Number 295845819).

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Ahmad, N. (2009). A Critical Appraisal of 'Triple Divorce' in Islamic Law. *International Journal of Law, Policy and the Family, 23*, 53–61.

Armstrong, D. (1971). Meaning and Communication. *The Philosophical Review, 80*, 427–447.

Austin, J. L. (1962). *How to Do Things with Words* (2nd ed.). Harvard University Press.

Bach, K. & R. Harnish (1979). *Linguistic Communication and Speech Acts*. MIT.

Damassino, N., & Novelli, N. (2020). Rethinking, Reworking, and Revolutionizing the Turing Test. *Minds and Machines, 30*, 463–468.
Ewing, J. (2017). *Faster, Higher, Farther: The Volkswagen Scandal.* Norton.
Freiman, O., & Miller, B. (2020). Can Artificial Entities Assert? In: S. Goldberg (Ed.), *Oxford Handbook of Assertion* (pp. 415–434). Oxford.
Goodfellow, I., Y. Bengio, & A. Courville (2016). *Deep Learning.* MIT.
Green, M. (2016). Assertion. In: D. Pritchard (ed.), *Oxford Handbooks Online*. Oxford.
Green, M. (2018). *Know Thyself: The Value and Limits of Self-Knowlege*. Routledge.
Green, M. (2019). Organic Meaning: An Approach to Communication with Minimal Appeal to Minds. In: A. Capone (ed.), *Further Advances in Pragmatics and Philosophy* (pp. 211–228). Springer.
Green, M. (2020). Speech Acts. In: E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/speech-acts/.
Green, M. (2021a). *The Philosophy of Language*. Oxford University Press.
Green, M. (2021b). Force, Content, and Translucent Self-Ascriptions. In: G. Mras & M. Schmitz (eds.), *Force, Content & the Unity of the Proposition* (pp. 195–214). Routledge.
Grice, H. P. (1957). Meaning. Reprinted in Grice 1989 (pp. 213–223).
Grice, H. P. (1975). Logic and Conversation. In: P. Cole & J. L. Morgan (eds.), *Syntax and Semantics, Vol. 3: Speech Acts* (pp. 41–58). New York.
Grice, H. P. (1989). *Studies in the Way of Words*. Harvard.
Kneer, M. (2021). Can a Robot Lie? Exploring the Folk Concept of Lying as Applied to Artificial Agents. *Cognitive Science*. https://doi.org/10.1111/cogs.13032
Krstic, V. (2018). Can You Lie without Intending to Deceive? *Pacific Philosophical Quarterly, 100*, 642–660.
Ludwig, K. (2020). Proxy Assertion. In: S. Goldberg (ed.), *Oxford Handbook of Assertion* (pp. 307–327). Oxford.
Malakoff, D. (2020). What if Biden Wins? *Science, 370*(6514), 284–285.
Maynard-Smith, J., & Harper, D. (2004). *Animal Signals*. Oxford University Press.
Michel, J. G. (2011). *Der qualitative Charakter bewusster Erlebnisse: Physikalismus und phänomenale Eigenschaften in der analytischen Philosophie des Geistes.* Brill/mentis.
Michel, J. G. (2019). How Are Species Discovered? Declarative Speech Acts in Biology. *Grazer Philosophische Studien, 96*(3), 419–441.
Michel, J. G. (2020). Could Machines Replace Human Scientists? Digitalization and Scientific Discoveries. In: B. Göcke & A. Rosenthal-von der Pütten (eds.), *Artificial Intelligence: Reflections in Philosophy, Theology, and the Social Sciences* (pp. 361–376). Brill/mentis.
Michel, J. G. (2022). Toward a Philosophy of Scientific Discovery. In: J. G. Michel (ed.), *Making Scientific Discoveries: Interdisciplinary Reflections* (pp. 9–53). Brill/mentis.
Nickel, P. (2013). Artificial Speech and Its Authors. *Minds and Machines, 23*, 489–502.
Reiter, E., & Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.
Roberts, C. (2018). Speech Acts in Discourse Context. In: D. Fogal, D. Harris, & M. Moss (eds.), *New Work on Speech Acts* (pp. 317–359). Oxford.
Sbisà, M. (2007). How to Read Austin. *Pragmatics, 17*(3), 461–473.
Searle, J. (1969). *Speech Acts*. Cambridge University Press.
Searle, J. (1983). *Intentionality*. Cambridge University Press.
Setiya, K. (2018). Intention. In: E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/fall2018/entries/intention/.
Stalnaker, R. (2014). *Context*. Oxford University Press.
Stainton, R. (2016). Full-on Stating. *Mind and Language, 31*(4), 395–413.
Strawson, P. F. (1964). Intention and Convention in Speech Acts. *The Philosophical Review, 73*(4), 439–460.