11 WHITE LIES ON SILVER TONGUES

WHY ROBOTS NEED TO DECEIVE (AND HOW)

Alistair M. C. Isaac and Will Bridewell

Deception is a regular feature of everyday human interaction. When speaking, people deceive by cloaking their beliefs or priorities in falsehoods. Of course, false speech includes malicious lies, but it also encompasses little white lies, subtle misdirections, and the literally false figures of speech that both punctuate and ease our moment-to-moment social interactions. We argue that much of this *technically* deceptive communication serves important pro-social functions and that genuinely social robots will need the capacity to participate in the human market of deceit. To this end, robots must not only recognize and respond effectively to deceptive speech but also generate deceptive messages of their own. We argue that deception-capable robots may stand on firm ethical ground, even when telling outright lies. Ethical lies are possible because the truth or falsity of deceptive speech is not the proper target of moral evaluation. Rather, the ethicality of human or robot communication must be assessed with respect to its underlying motive.

The social importance of deception is a theme that emerges repeatedly in fictional portrayals of human–robot interaction. One common plot explores the dangers to society if socially engaged robots lack the ability to detect and respond strategically to deception. For instance, the films *Short Circuit 2* (1988), *Robot & Frank* (2012), and *Chappie* (2015) all depict naive robots misled into thievery by duplicitous humans. When robots are themselves deceivers, the scenarios take on a more ominous tone, especially if human lives are at stake. In *Alien* (1979) the android Ash unreflectively engages in a pattern of deception mandated by its owners' secret instructions. Ash's inhuman commitment to the mission leads to the grotesque slaughter of its crewmates. Regardless of whether Ash can recognize its actions as deceptive, it can neither evaluate nor resist the ensuing pattern of behavior, which has dire consequences for human life. *2001: A Space Odyssey* (1968) provides a subtler cinematic example in the computer HAL. HAL's blind adherence to commands to deceive, unlike Ash's, is not what results in its murder of the human crew. Instead, the fault rests in the computer's inability to respond strategically—humanely—to the mandate that the goal of the mission be kept secret. When the demands of secrecy require HAL to lie in subtle ways, it instead turns to murder. Ironically, it is HAL's *inability to lie effectively* that leads to catastrophe.

More lighthearted depictions of social robots have found comedic traction in the idea that the ability to deceive is a defining feature of humanity. For instance, in the TV comedy *Red Dwarf* (1991), Lister, a human, teaches the robot Kryten to lie in an attempt to liberate it from the inhuman aspects of its programming. Conversely, Marvin ("the Paranoid Android") in *The Hitchhiker's Guide to the Galaxy* (Adams 1980) is characterized by its proclivity to tell the truth, even when grossly socially inappropriate. Marvin's gaffs may be funny, but they underscore the importance of subtle falsehoods in upholding social decorum. Furthermore, the most human (and heroic) fictionalized robots display a versatile capacity to mislead. *Star Wars*' (1977) R2D2, for instance, repeatedly deceives the people and robots around it, both subtly through omission and explicitly through outright lies, in service to the larger goal of conveying the plans for the Death Star to rebel forces. This pattern of deception is part of what gives a robot that looks like a trashcan on wheels the human element that endears it to audiences.

In the remainder of the chapter, we develop an account of how, why, and when robots should deceive. We set the stage by describing some prominent categories of duplicitous statements, emphasizing the importance of correct categorization for responding strategically to deceptive speech. We then show that the concept of an *ulterior motive* unifies these disparate types of deception, distinguishing them from false but not duplicitous speech acts. Next, we examine the importance of generating deceptive speech for maintaining a pro-social atmosphere. That argument culminates in a concrete engineering example where deception may greatly improve a robot's ability to serve human needs. We conclude with a discussion of the ethical standards for deceptive robots.

11.1 Representing Deception

We care if we are deceived. But what does it mean to be deceived, and why do we care about it? Intuitively, there is something "bad" about deception, but what is the source of that "badness"? This section argues that we are troubled by deceit because it is impelled by a covert goal, an *ulterior motive*. We need to detect deception because only by identifying the goals of other agents, whether human or robot, can we respond strategically to their actions. This insight highlights a specific challenge for savvy agents: effective detection of deception requires inferences about the hidden motives of others.

11.1.1 Taxonomy of Deception

The following scenario steps away from the world of robots to tease apart the components of deception. In this interaction, Fred is a police detective investigating a crime; Joe is a suspect in the crime; and Sue is Joe's manager at work.

FRED: Where was Joe on the morning of March 23rd?
SUE: Joe was here at the office, working at his desk.

Consider the most straightforward form of deceptive speech: *lying*. At a first pass, lying occurs when an agent willfully utters a claim that contradicts his beliefs. The detective suspects Joe of having committed a crime on the morning of March 23rd, so he cares where Joe happened to be. Is Sue lying about Joe's location? The answer is important not only because Fred cares about Joe's location but also because he cares about Joe's accomplices. If Sue is lying to protect Joe, she may be implicated in the crime.

How can Fred tell whether Sue is lying? Suppose, for instance, the detective has access to security video from the morning of March 23rd at the scene of the crime—that is, he knows that Sue's claim that Joe was at the office is false. Is this fact enough to determine that Sue is lying? Not necessarily; for instance, Sue might only have a *false belief*. Perhaps she saw a person who looks like Joe at the office and formed the erroneous belief that Joe was at work. Similarly, Sue might be *ignorant* of Joe's actual whereabouts. Maybe she has no evidence one way or another about Joe's presence, and her response is a rote report on assigned employee activities for March 23rd.

The falsity of a statement is not, then, sufficient evidence of lying. However, combined with other evidence, for instance biometric cues such as excessive sweat, fidgeting, or failure to make eye contact, it may nevertheless allow Fred to confidently infer that Sue is indeed lying. But is determining correctly whether or not Sue is lying enough for Fred's purposes? Consider two possibilities: in one scenario, Sue lies because she is in cahoots with Joe and has received a cut of the loot; in the other, Sue lies because she was hungover the morning of March 23rd, arrived late, and is scared of a reprimand for her unexcused absence. In both situations Sue is lying, but the appropriate response by the detective is radically different. In the first case, Fred might charge Sue with abetting a crime; in the second, he may simply discount Sue's testimony.

The point of the example is twofold. First, detecting a lie requires knowledge about more than a single statement's accuracy. A lie depends constitutively on the speaker's state of mind, so her intent to deceive must be inferred. Second, the correct response to a lie may turn on more than the mere fact that a lie has been uttered. Crucially, the responder's strategy may depend on the goal that motivated the lie—it is this goal that underpins any intent to deceive.[1] These two key features characterize other forms of deception identified in the literature. Here we briefly consider paltering, bullshit, and pandering (for a more in-depth treatment, see Isaac and Bridewell 2014).

Paltering (Schauer and Zeckhauser 2009; Rogers et al. 2014) occurs when a speaker misleads his interlocutor by uttering an irrelevant truth. A paradigmatic example is the used-car salesman who truthfully claims, "The wheels on this car are as good as new," to direct attention away from the poor quality of the engine. Paltering illustrates that neither the truth-value of an utterance nor the speaker's belief about its truth are crucial factors for determining whether the utterance is deceptive. Rather, the ethical status of speech may turn entirely on whether it is motivated by a malicious intent to misdirect.

Bullshit (Frankfurt 1986; Hardcastle and Reisch 2006) occurs when a speaker neither knows nor cares about the truth-value of her utterance. Bullshit may be relatively benign, as in a "bull session" or the exchanging of pleasantries around the water cooler. However, there is malicious bullshit as well. A confidence man may spew bullshit about his background and skills, but if people around him believe it, the consequences may be disastrous. Frank Abagnale, Jr., for instance, repeatedly impersonated an airline pilot to travel for free (events dramatized in the 2002 film *Catch Me If You Can*), but his bullshit put lives at risk when he was asked to actually fly a plane and blithely accepted the controls.

Pandering is a particularly noteworthy form of bullshit (Sullivan 1997; Isaac and Bridewell 2014). When someone panders, he (may) neither know nor care about the truth of his utterance (hence, a form of bullshit), but he does care about an audience's perception of its truth. A politician who, when stumping in Vermont, proclaims, "Vermont has the most beautiful trees on God's green earth!" does so not because she believes the local trees are beautiful, but because she believes the local audience believes Vermont's trees are beautiful—or, more subtly, that the locals want visitors to believe their trees are beautiful.

Lying, paltering, bullshitting, and pandering are all forms of deception. However, they are united neither by the truth-value of the utterance nor the speaker's belief in that utterance. Moreover, bullshitting and pandering may lack even an intent to deceive. Rather, what unites these categories of perfidy is the presence of a goal that supersedes the conversational norm for truthful speech. The nature of this goal, in addition to reliable detection and classification of deception, is vital information for any agent forming a strategic response.

11.1.2 Theory of Mind, Standing Norms, and Ulterior Motives

What capacities does a robot require to identify and respond to the wide variety of deceptive speech? An answer to this question is undoubtedly more complex than one we can currently provide, but a key component is the ability to represent the mental states of other agents. Socially sophisticated robots will need to track the beliefs and goals of their conversational partners. In addition, robots' representations will need to distinguish between baseline goals that may be expected of any social agent, which we call *standing norms*, and the special goals that supersede them, which we call *ulterior motives*.

When we claim that deception-sensitive robots will need to track the beliefs and goals of multiple agents, we are stating that these robots will need a *theory of mind*. This phrase refers to the ability to represent not only one's own beliefs and goals but also the beliefs and goals of others. These representations of the world may conflict, so they must be kept distinct. Otherwise, one could not tell whether one believed that lemurs make great pets or one believed that someone else believed it. As an illustration, suppose that Sue and Joe, from the earlier example, are accomplices. In that case, Sue believes that Joe was at the scene of the crime but wants Frank to believe that Joe was at work. If she thinks her lie was compelling, then Sue will form the belief that Frank believes that Joe was at the office—this is a first-order theory of mind. Pandering requires a second-order theory of mind. For instance, the politician must represent his (zeroth-order) belief that the audience (first order) believes that he (second order) believes the trees in Vermont are beautiful (figure 11.1).

Given a theory of mind rich enough to represent the beliefs and goals of other agents several levels deep, what else would a robot need to strategically respond to deception? According to our account, socially aware robots would need to represent the covert motive that distinguishes deception from normal speech. Fortunately, even though the particular motive of each deceptive act generally differs (e.g., Sue's goal in lying differs from that of the used-car salesman in paltering), there is a common factor: a covert goal that trumps expected standards of communication. Therefore, to represent a deceptive motive, a robot must distinguish two types of goals: the typical and the superseding.

We call the first type of goal a *standing norm* (Bridewell and Bello 2014), a persistent goal that directs an agent's typical behavior. For speech, Paul Grice introduced a set of *conversational maxims* (1975) that correspond to this notion of a standing norm. His maxim of quality, frequently glossed as "be truthful," is the most relevant to our discussion. Grice argued that people expect that these maxims will be followed during ordinary communication and that flagrant violations cue that contextual influences are modifying literal meaning. The crucial point for our purposes is that *be truthful* is a goal that plausibly operates under

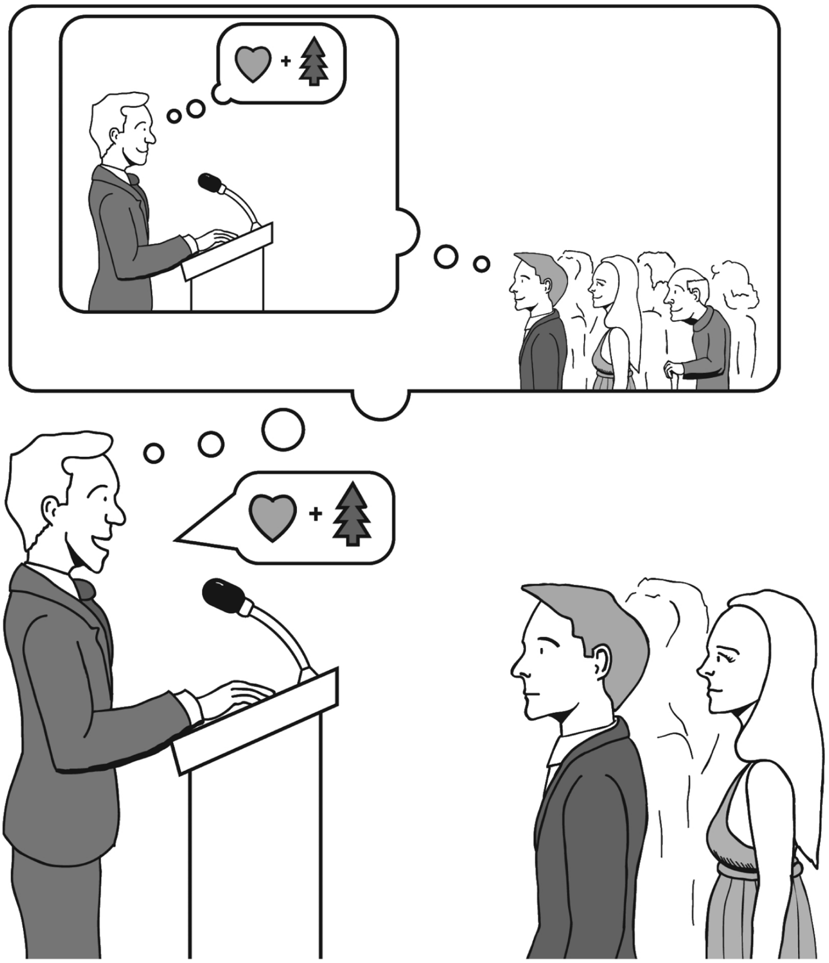

FIGURE 11.1. Pandering requires a second-order theory of mind. The successful panderer believes (zeroth order) that the listener believes (first order) that the speaker believes (second order) his utterance. Image credit: Matthew E. Isaac.

all typical circumstances. Other standing norms might regulate typical speech in subtler ways (e.g., *be polite* or *be informative*).

In any complex social situation, more than one goal will be relevant. If these goals suggest conflicting actions, we will need some method to pick which goals to satisfy and which to violate. For instance, imagine a friend who composes saccharine poems to his dog asks your opinion on the latest one. The conflict between the standing norms *be truthful* and *be polite* may likely pose a dilemma. If you are lucky, you may get away with satisfying both: "I can tell that you put a lot of work into that poem. You must really love your corgi." This response is both truthful and polite, yet it achieves this end by avoiding a direct answer to your friend's question. Notice the difference between this kind of misdirection and paltering, however; we don't typically think of an answer such as this as deceptive because there is no hidden goal—it is governed entirely by the expected standards of conversation.

If your friend presses you, demanding an explicit opinion on his poetry, you will need some heuristic to determine which of your standing norms to violate. One way to achieve this end is to prioritize your goals. Several factors—situational, cultural, emotional—will determine which norm takes priority and guides speech. Ranking truth over civility may lead to a brutal dismissal of your friend's literary skills. Ranking civility over truth may lead to false praise. Such false praise is *technically* deceitful—we intend our friend to form a belief that conflicts with the truth.

We often categorize an utterance such as this, the false praise of a friend's inept poem, as a *white lie*. On the one hand, we recognize that it is technically an act of deceit, because a goal has superseded the norm *be truthful*, and in this sense a "lie." On the other hand, since the superseding goal is itself a standing norm (in this case *be polite*), it bears no malicious intent, and we do not typically judge the lie to be morally reprehensible. The situation is different when the superseding goal is not a standing norm. In that case, we refer to the goal prioritized over a norm as an *ulterior motive*. The presence of a relevant ulterior motive differentiates a maliciously deceptive utterance from a benign one. If you praise your friend's poetry, not because you prioritize the norm *be polite*, but because a goal to borrow money from your friend supersedes all your standing norms, then we would no longer judge your false praise to be morally neutral. Revisiting Sue, if her false response to the detective is grounded in a false belief, she is not suppressing a conversational norm and her error is innocent. However, if Sue has a goal to *protect Joe* that supersedes the conversational norm *be truthful*, then her response is deceptive. Other ulterior motives in the earlier examples include *sell this car* and *get elected*.

To summarize, if a robot is to effectively recognize deceptive speech and respond strategically to it, it must be able to represent (a) the difference between its own beliefs and desires and those of its interlocutor; and (b) the difference between standing norms of behavior and ulterior motives. Of course, other capacities are also needed, but we claim that they will appeal to these representations. We next turn to the question of the "badness" of deception and argue that some forms of deception are desirable, even in robots.

11.2 Deceiving for the Greater Good

So far, we have seen that robots must possess a theory of mind in order to respond effectively to deceptive communication and, in particular, the ability to identify ulterior motives. But an effective social robot cannot treat all deceptive speech as malign, or all agents who act on ulterior motives as malicious. Furthermore, such a robot may find itself following ulterior motives, and its (technically) deceptive behavior may have a positive social function. After discussing examples of the pro-social function of deceptive speech, we consider some cases where we want robots to lie to us.

11.2.1 Benign Deceptions

When we talk to each other, our words do far more than communicate literal meaning. For instance, we routinely exchange pleasantries with co-workers. Two people walk toward each other in a hallway, one asks how the other's day is going, and the response is a casual "fine." This sort of exchange reinforces social-group membership. The colleagues recognize each other in a friendly way and the literal content serves only a secondary function, if any. Other examples include "water cooler" conversations about the weather or sports, or office gossip.

Often these casual conversations are forms of bullshit: conversants may neither know nor care about the truth, but the conversation goes on. Speculating who will win the World Cup or whether there will be rain on Sunday seems generally unimportant, but people talk about these topics routinely. In addition, consider all the times that people exchange pleasantries using outright lies. For instance, we might compliment a friend's trendy new hairstyle, even if we think it is hideous. In these cases, affirming the value of peers can take priority over conveying truth or engaging in debate. In fact, treating such pleasantries as substantive may lead to confusion and social tension. Responding to a co-worker's polite "Hi, how are you?" with "My shoulder aches, and I'm a bit depressed about my mortgage" will, at the least, give the colleague a pause. Continued tone-deaf responses will likely reduce the opportunities for reply. Treating a rote exchange of pleasantries as legitimately communicative not only is awkward, but undermines the exchange's pro-social function (Nagel 1998).

Another common form of benignly deceptive speech includes metaphors and hyperbole: "I could eat a horse"; "These shoes are killing me"; or even "Juliet is the sun. Arise, fair sun, and kill the envious moon." There is nothing malevolent about these figures of speech, but to respond to them effectively requires an ability to recognize the gap between their literal content and the speaker's beliefs. If a chef served a full-sized roast equine to a customer who announced, "I could eat a horse," she would be greeted with surprise, not approbation. Exactly how to compute the meaning of metaphorical and hyperbolic expressions is a vexed question, but we can agree that these figures of speech violate standing norms of the kind articulated by Grice (1975; Wilson and Sperber 1981), and thus technically satisfy the earlier definition of deceptive speech.

It would be fair to ask whether metaphorical speech is necessary for effective communication. What does a turn of phrase add? Are there some meanings that can be conveyed only through metaphor, having no paraphrase into strictly literal language (Camp 2006)? Even if the answer is negative, metaphor and hyperbole provide emphasis and add variety, color, and emotion to conversations. We would describe someone who avoids figures of speech or routinely misunderstands them as difficult to talk to, socially awkward, or even "robotic."

More complex social goals may also require systematic, arguably benign deception; consider, for instance, *impression management*. In complex social interactions, what others think about us, their impression of our character, is often important. How we appear to people in power or to our peers has very real effects on our ability to achieve long-term goals and maintain social standing. Managing these impressions to achieve broader goals may supersede norms of truthfulness in conversation. For instance, suppose a worker wants his boss to see him as loyal to her. To this end, he supports her attempts to close a small deal with a corporate ally even though he disagrees with the content. His goal to manage his appearance as a loyal employee motivates him to vote in favor of his boss's deal at the next meeting.

In this example, the long-term goal of demonstrating political solidarity swamps short-term concerns about relatively unimportant decisions. Moreover, disingenuously endorsing less important proposals in the short term may give the employee the cachet to speak honestly about important deals in the future. Whether one thinks the subtle politics of impression management are morally permissible, they certainly play a major role in the complex give-and-take characteristic of any shared social activity, from grocery shopping with one's family to running a nation-state. As we will see, simple impression management may be required for practical robot applications in engineering.

11.2.2 When We Want Robots to Lie

Do we want robots that can banter about the weather and sports, compliment us on a questionable new haircut, or generate appropriately hyperbolic and metaphorical expressions? ("This battery will be the death of me!") If we want robots that can smoothly participate in standard human modes of communication and social interaction, then the answer must be *yes*. Even if our primary concern is not with fully socially integrated robots, there are many specific practical applications for robot deception. Here we consider only two: the use of bullshit for self-preservation and the use of systematic lies to manage expectations about engineering tasks.

11.2.2.1 Bullshit as Camouflage

If a primary function of bullshit is to help one fit into a social group, that function may be most important when one is not in fact a bona fide member of the group in question. Consider, for instance, the case of a computer scientist at a sports bar. For him, the ability to bullshit about sports, to make the right kinds of comments even if he neither knows nor cares about their truth, can mean the difference between treatment as a peer and humiliation or physical assault. In situations like this, the ability to spew the right sort of bullshit acts as a form of camouflage, enabling the computer nerd to superficially fit in with the people around him.

This sort of camouflage is not always benign, but the skill may be vital to survival for some kinds of robots. The spy, the fifth columnist, and the terrorist all use this technique, not because it is inherently nefarious, but because the ability to fit in can be critical for self-preservation. A robot working in a hostile community or confronted by belligerent locals in the wrong part of town would benefit from the ability to bullshit judiciously, whether the hostiles themselves were human or mechanical. A socially sophisticated robot should be able to generate appropriate signals to blend in with any conversational community and thereby extricate itself from dangerous situations without violence or social friction.[2]

In general, we should acknowledge that robots will inevitably be seen as part of an out-group, as less than human, regardless of their social prowess. Thus, their ability to bullshit their way into our hearts and minds may be key to their broad acceptance into the workforce. As objects whose physical well-being may constantly be under threat, robots might use bullshit effectively to provide a verbal buffer against prejudice.

11.2.2.2 Managing Impressions to Manage Uncertainty

Our earlier example of impression management was political in flavor, but the basic concept has more mundane applications as well. In engineering, there is a common practice of knowingly overstating the time it will take to complete a task by a factor of two or three. This convention is sometimes referred to as the *Scotty Principle* after the character in Star Trek.[3] This category of lying serves two major functions. First, if the engineer finishes ahead of time, she looks especially efficient for having delivered results sooner than expected (or she rests a bit and reports an on-time completion). Second, and more important for our argument, the engineer's estimate creates a clandestine buffer for contingencies that protects her supervisor from making aggressive, time-sensitive plans.

At a practical level, the Scotty Principle factors in the completely unexpected: not just any initial failure at assessing the intrinsic difficulty of the task, but also unforeseen extrinsic factors that might impact successful completion (a strike at the part supplier's warehouse, a hurricane-induced power outage, etc.). Such "unknown unknowns" (those facts that we do not know that we do not know) famously pose the greatest challenge to strategic planning. Unlike known unknowns, which can be analyzed quantitatively and reported as confidence intervals or "error bars," unknown unknowns resist any (meaningful) prior analysis—we cannot plan for a difficulty we do not expect or even imagine. Yet inserting a temporal buffer between the known aspects of a repair job and those that are unknown does in some way prepare us for the unexpected. Furthermore, the practice acts as a deliberate corrective to the engineer's own potential failings, including any lack of self-knowledge about her skills and speed of work. Ironically, then, *willfully deceiving a supervisor may correct for an engineer's self-deception.*

This example was not picked arbitrarily. Engineering tasks, including the repair of technical equipment or software systems, are potential applications for sophisticated robots. Do we want robotic engineers to lie to us systematically in exactly this way? Plausibly, the answer is yes: if we want our best robot engineers to meet the standards of our best human engineers, then we should expect them to embrace the Scotty Principle. Nevertheless, we have encountered two prominent objections to this line of reasoning, which we consider in turn.

The first objection is that current human-engineering practice sets too low a bar for assessing future robot engineers. We should expect to be able to improve our robots until they *correctly* estimate their capabilities and thereby avoid the need for recourse to Scotty's Principle. Yet this idealized view of robot performance belies the nature of engineering and the capacity for self-knowledge in dynamic, complex environments. Contingencies arise in engineering, and novel tasks stretch the scope of an engineer's abilities, whether human or mechanical. Aiming for a robot that predicts the results of its interventions in the world with perfect accuracy is to aim for the impossible. Reliable engineering practice requires accepting that unknown unknowns exist and preparing for the possibility that they may confront a project when it is most inconvenient.

The second, distinct worry arises for those who acknowledge the danger of unknown unknowns, yet insist it is safer to leave contingency planning in the hands of human supervisors than to permit robot engineers to systematically lie. But this suggestion relies on an unrealistic assessment of human nature—one that may be dangerous. Anyone who has worked with a contractor either professionally or personally (e.g., when remodeling or repairing a house) will be familiar with the irresistible tendency to take at face value that contractor's predictions about job duration and cost. If these predictions are not met, we hold the contractor accountable, even if he has been perpetually tardy or over-budget in the past—we certainly do not blithely acknowledge he may have faced unforeseen contingencies. Yet this tendency is plausibly even greater in our interactions with robots, which we often assume are more "perfect" at mechanical tasks than in fact they are. Paradoxically, if we insist all contingency planning must reside with human supervisors, we will need to train them *not* to trust their robot helpers' honest predictions of task difficulty!

The apparent infallibility of robots is exactly why we must ensure that socially sophisticated robots can misrepresent their predictions about the ease of a job. This added step can correct for both the robot's self-deception about its abilities and its human user's unrealistic optimism about robot dependability.

11.3 Ethical Standards for Deceptive Robots

We have just suggested that robots will need the capacity to deceive if they are to integrate into and contribute effectively to human society. Historically, however, many ethical systems have taught that lying and other forms of deception are intrinsically wrong (e.g., Augustine 1952). If we aim for deception-capable robots, are we giving up on the possibility of ethical robots? We argue that the answer is a resounding *no*. This is because the locus of ethical assessment is not in the content of speech, but in the ulterior motive that gives rise to it. Therefore, ensuring that a social robot behaves morally means ensuring that it implements ethical standing norms and ranks them appropriately.

Precisely how is it possible for deception to be ethical? In section 11.1 we argued that the unifying factor for all forms of deceptive speech is the presence of an ulterior motive, a goal that supersedes standing norms such as *be truthful*. Paradigmatic cases of morally impermissible ulterior motives are those involving pernicious goals, such as the concealment of a crime or the implementation of a confidence scheme. In contrast, section 11.2 surveyed examples where deceptive speech serves pro-social functions and the ulterior motives are pro-social goals, such as boosting a colleague's self-esteem or establishing trust. There is a discriminating factor between right and wrong in these cases; however, it depends not on the deceptiveness of the speech per se but on the goals that motivate that speech.

If this analysis is correct, it implies that an ethical, deception-capable robot will need the ability to represent and evaluate ranked sets of goals. To draw an example from literature, consider Isaac Asimov's (1950) Three Laws of Robotics, an early proposal for a system of robot ethics. Asimov's "laws" are really inviolable, prioritized, high-level goals or, in our terminology, a ranked set of standing norms combined with the stipulation that no ulterior motives may supersede them. Notably, accepting this system means accepting deception by robots, because Asimov's stipulation ensures that his laws will be ranked above the norm *be truthful*. For instance, the First Law is that "a robot may not injure a human being or, through inaction, allow a human being to come to harm," and the Second Law is that "a robot must obey the orders given it by human beings except where such orders would conflict with the First Law." Suppose a murderer comes to the door of a home and orders a robot butler to tell him where his intended victim is hiding. Here, the Second Law mandates that the robot answer, but to satisfy the First Law it must lie to protect the victim's life.

Asimov's system is based on a set of duties or obligations, which makes it a deontological ethical theory (Abney 2012). Traditionally, this approach has produced the most severe criticisms of the permissibility of lying, so perhaps it is comforting to see that deceptive robots may conform to a deontological ethics. Immanuel Kant, the most famous deontologist, repeatedly and passionately defended the extreme position that deceptive speech is never permissible *under any circumstance*. He even addressed the foregoing scenario, where a murderer asks someone where to find a would-be victim. Kant (1996) concluded that even to save a life, lying is impermissible. Most ethicists have found this conclusion distasteful, and nuanced discussions of the permissibility of lying often proceed by first categorizing lies in terms of the respective ulterior motives that produce them, then assessing whether these motives should indeed be allowed to supplant the mandate for truthfulness (e.g., Bok 1978).

The perspective outlined here is also compatible with consequentialism, the main alternative to deontological theories. Consequentialists assess the morality of an action on the basis of its consequences—good acts are those that produce more of some intrinsically good property in the world, such as well-being or happiness, while bad acts are those that produce less of it. How does this view evaluate deceptive speech? If the speech has overall positive consequences, increasing well-being or happiness, then whether or not it is deceptive, it is permissible, perhaps even mandatory. The influential consequentialist John Stuart Mill addressed this topic: while placing a high value on trustworthiness, he nevertheless asserted the permissibility to deceive "when the withholding of some fact . . . would save an individual . . . from great and unmerited evil, and when the withholding can only be effected by denial" (1863, ch. 2).

By shifting the locus of moral assessment to the speaker's ulterior motives, we have not made the problems of robot ethics any more complex; however, we have also not made the problems any simpler. A deontological specification of duties or a consequentialist calculation of overall well-being remains equally challenging regardless of whether the robot may deceive. The computational deontologist remains burdened with questions about the relative priorities of norms or goals, along with other general challenges related to formalizing ethical maxims and using them to make inferences about an action's morality (Powers 2006). Likewise, the computational consequentialist still faces the challenge of determining and comparing the effects of potential actions (whether deceptive or not) in any given situation. On this point, Keith Abney argues that a simple-minded consequentialism "makes moral evaluation impossible, as even the short-term consequences of most actions are impossible to accurately forecast and weigh" (2012, 44).

To conclude, our basic proposal is that effective social robots will need the ability to deceive in pro-social ways, an ability that may facilitate the integration of android assistants into society, preserve robot safety in the face of prejudice, and protect humans from our own misconceptions about the infallibility of our mechanical helpmates. When assessing the ethicality of speech, the proper target for evaluation is the motivating goal, not the truth or falsity of the speech per se. Consequently, permitting robots to lie does not substantively change the technical challenges of ensuring they behave ethically. Nevertheless, there are challenges distinctive to the design of a deception-capable robot, as it requires a theory of mind and, in particular, the capacity to detect and reason about ulterior motives.

Acknowledgments

We thank Paul Bello for discussions that helped inform the text. Will Bridewell was supported by the Office of Naval Research under grant N0001416WX00384. The views expressed in this chapter are solely the authors' and should not be taken to reflect any official policy or position of the U.S. government or the Department of Defense.

Notes

1. There is a technical literature on how best to define *lying*, and one of the most debated issues is whether an "intent to deceive" need be present (for a survey, see Mahon 2016). On our view, an "intent to deceive" is just one possible instance of (or consequence of) an *ulterior motive*; this analysis both avoids common counterexamples to the "intent" condition and unifies the definition of lying with that of other forms of deception.
2. The crude beginnings of this challenge are already with us. For instance, self-driving cars must find a way to blend in on roads dominated by human drivers, and an accident or impasse may result from their blind adherence to the letter of traffic law (veridicality) and inability to interpret or send the subtle road signals required to fit in with the rest of traffic (bullshit). A classic example is the four-way stop sign, where self-driving cars have become paralyzed when none of the other cars come to a complete stop. Effective navigation of such intersections requires coordinating behavior through nuanced signals of movement, sometimes even bluffs, rather than strict deference to the rules of right-of-way.
3. Throughout the *Star Trek* TV series, engineer Scotty routinely performs repairs faster than his reported initial estimates. This phenomenon, and the Scotty Principle itself, is explicitly acknowledged in the film *Star Trek III: The Search for Spock* (1984):

   KIRK: How much refit time until we can take her out again?
   SCOTTY: Eight weeks sir, but you don't have eight weeks, so I'll do it for you in two.

KIRK: Mr. Scott, have you always multiplied your repair estimates by a factor of four?

SCOTTY: Certainly sir, how else can I keep my reputation as a miracle worker?

Works Cited

Abney, Keith. 2012. "Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George A. Bekey, 35–52. Cambridge, MA: MIT Press.

Adams, Douglas. 1980. *The Hitchhiker's Guide to the Galaxy*. New York: Harmony Books.

Asimov, Isaac. 1950. *I, Robot.* New York: Doubleday.

Augustine. (395) 1952. "Lying." In *The Fathers of the Church* (vol. 16: *Saint Augustine Treatises on Various Subjects*), edited by Roy J. Deferreri, 53–112. Reprint, Washington, DC: Catholic University of America Press.

Bok, Sissela. 1978. *Lying: Moral Choice in Public and Private Life*. New York: Pantheon Books.

Bridewell, Will and Paul F. Bello. 2014. "Reasoning about Belief Revision to Change Minds: A Challenge for Cognitive Systems." *Advances in Cognitive Systems* 3: 107–22.

Camp, Elisabeth. 2006. "Metaphor and That Certain 'Je ne Sais Quoi.'" *Philosophical Studies* 129: 1–25.

Frankfurt, Harry. (1986) 2005. *On Bullshit*. Princeton, NJ: Princeton University Press.

Grice, H. Paul. 1975. "Logic and Conversation." In *Syntax and Sematics 3: Speech Acts*, edited by Peter Cole and Jerry L. Morgan, 41–58. New York: Academic Press.

Hardcastle, Gary L. and George A. Reisch. 2006. *Bullshit and Philosophy*. Chicago: Open Court.

Isaac, Alistair M. C. and Will Bridewell. 2014. "Mindreading Deception in Dialog." *Cognitive Systems Research* 28: 12–9.

Kant, Immanuel. (1797) 1996. "On a Supposed Right to Lie from Philanthropy." In *Practical Philosophy*, edited by Mary J. Gregor, 605–15. Reprint, Cambridge: Cambridge University Press.

Mahon, James Edwin. 2016. "The Definition of Lying and Deception." In *The Stanford Encyclopedia of Philosophy*, Spring 2016 ed., edited by Ed N. Zalta. http://plato.stanford.edu/archives/spr2016/entries/lying-definition/.

Mill, John Stuart. 1863. *Utilitarianism*. London: Parker, Son & Bourn.

Nagel, Thomas. 1998. "Concealment and Exposure." *Philosophy and Public Affairs* 27: 3–30.

Powers, Thomas M. 2006. "Prospects for a Kantian Machine." *IEEE Intelligent Systems* 21: 46–51.

Rogers, Todd, Richard J. Zeckhauser, Francesca Gino, Maurice E. Schweitzer, and Michael I. Norton. 2014. "Artful Paltering: The Risks and Rewards of Using

Truthful Statements to Mislead Others." HKS Working Paper RWP14-045. Harvard University, John F. Kennedy School of Government.

Schauer, Frederick and Richard J. Zeckhauser. 2009. "Paltering." In *Deception: From Ancient Empires to Internet Dating*, edited by Brooke Harrington, 38–54. Stanford, CA: Stanford University Press.

Sullivan, Timothy. 1997. "Pandering." *Journal of Thought* 32: 75–84.

Wilson, Deirdre and Dan Sperber. 1981. "On Grice's Theory of Conversation." In *Conversation and Discourse*, edited by Paul Werth, 155–78. London: Croom Helm.