Why Robotic Ethical Claims Will Outpace Their Understanding

Kevin Gold

Yale University, New Haven, CT 06511

kevin.gold@yale.edu

Nico: a robot that says "I"

Recently, I added some code to the Yale Social Robotics Lab's robot Nico that allows it to learn some very interesting words: "I" and "you." When I say that Nico learned these words, that is bound to cause some confusion and controversy, so what I mean is that Nico had these words in its speech recognition lexicon, but it had no particular properties or definitions associated with the words. By observing a game of "catch," the robot learned to associate these words with the conversational roles of "speaker" and "addressee." That is to say, when the robot heard the phrase "I got the ball" or "you got the ball," it would look to see who had the ball, associate the relevant word with the all the properties of that person, and eventually use the weight of statistical evidence to infer that "I" and "you" referred to speaker and addressee, respectively. In doing so, it ruled out other properties that did not shift with conversational role (Gold and Scassellati, 2006b).

The state of Nico's knowledge after this experiment was such that it would be able to understand who was being talked about when it heard statements using the words "I" and "you," so long as it could tell who was speaking and who was being addressed. To use allow the robot to use the words itself, however, required a bit more work. Specifically, the robot had to think of itself as a person.

Again, *think* is something of a colloquialism here, of the kind that is practically unavoidable when dealing with intelligent agents. What I mean is that in its internal world model, the robot had to represent the fact that as it conveyed a message, it was taking on the role of "speaker," and the person to whom it communicated was taking on the role of "addressee." I could see no way for the robot to learn this through observation. It is not at all obvious to a robot that when it communicates, it is fulfilling the same role as a human that speaks. This is especially true because at the time of this writing, Nico still produces messages only through on-screen text – so even the modality of its speech is different from its observations. In the end, Nico had to be programmed with an assumption before it could use "I": when it communicated, it was assuming the role of "speaker,"

which was exactly the same role as when a human communicated. It could then use its learned knowledge that "I" corresponded to this property to make the statement, "I got the ball" when it was close to its ball (Gold and Scassellati, 2006a).

I predict that this kind of assumed human-equivalence will be useful enough that it will become ubiquitous in robots that can learn to use language on the fly. Language is so tailored to human experience that it will be far more convenient for robots to use existing metaphors and concepts to refer to themselves than to try to invent new language for them. With physical terms, this is fairly uncontroversial. A robot doesn't robot-trip and robot-fall; it trips and falls, even if its structure is different. When it comes to mental language, the same liberal attitude will still apply, so long as the robot's behavior is analogous to the human case. A robot won't be said to robot-believe, robot-hope, or become robot-sad; people will say that a robot believes, hopes, or is sad.

Robots that learn language will learn to apply these terms to themselves, whether or not humans do so. The world is (and will be) much more full of humans than language-using robots; a robot that does not use every opportunity to co-opt language will not be very good at it. But this may lead to an unintended consequences, which people will undoubtedly find unnerving at first. Once a robot can learn language in general, there is nothing preventing it from learning to make ethical claims about itself. Its assumption of human equivalence will not have come from its own reasoning; it will have come, by necessity, from humans.

Assumed Human Equivalence

Having sketched out the path of my argument, I will describe some of the reasoning above in more detail before delving into its implications. A reader interested in only the ethical implications can skip to the next section.

First, I make the claim that we will want to have robots that can learn new words and terminology. If this were not the case, it would not necessarily be true that robots would freely use human-centric language to refer to themselves. A

robot that has all of its vocabulary preprogrammed can have any number of other workarounds. A robot designer may provide the robot with two separate definitions for many terms – one that encompasses the robot meaning, and another that encompasses the human meaning. The roboticist may omit certain problematic words, especially in relation to the robot; the language of ethics, for example, will probably be one of the last things to ever cross a designer's mind in designing a task-specific robot. Essentially, if the designer has strict control over a robot's language, the robot will use language in the manner specified by the designer, and that is that.

This could work for a long time, particularly if language-using robots are only used in a small number of applications in a small number of languages. But eventually, somebody will realize that this approach simply doesn't scale well. Every possible endeavor in which a robot could find application, from home repair to search and rescue, has its own technical jargon, an evolving slang for that jargon, and a wealth of circumstances that simply can't be predicted by a designer. A language-enabled bomb squad robot told "It's under the ottoman" will probably be at a loss if the robot's designer hand-coded a vocabulary specific to bomb retrieval. Coding huge vocabularies in other languages will likewise seem something of a wasted effort; we know that humans regularly solve the problem of learning an arbitrary subset of an arbitrary language, so there is no reason to treat this problem as insoluble.

It may be the case that once a single robot learns a certain amount of vocabulary, its memory would be copied and distributed to other robots, perhaps with the learning disabled. This would be a practical way of ensuring that every robot does not need to relearn basic vocabulary. But if an algorithm is already known for learning language, there would be no practical reason for disabling it in the robot copies unless the algorithm were particularly unstable. The robots may as well continue learning on the job.

Thus, there are good reasons to believe that robots that can learn new words will be seen as eminently practical. The next link in the argument, and the crucial one, is that to pragmatically learn language, robots must make some kind of assumption of human equivalence.

This is not to say that robots must assume that they are equal in *value* to humans merely to learn language. It is possible for a robot to learn some language without ever dealing with such questions at all, or without even having a way to represent such statements in their data structures. Nico's world, for example, consists of a two-dimensional map of where people are, their current conversational roles, their identities, the location of its ball, and a variable indicating who the ball is closest to. Nico currently has no way of learning any words that do not refer to one of these properties.

But to the extent that Nico represents itself in memory at all, it does so using the same data structure format as the other people in the scene: location, conversational role, and whether it has the ball. If Nico's representation of others were to become more complex, its representation of itself would become more complex as well, because it is simply easiest to program Nico to treat itself as a special case of a person. Doing so allows Nico to use any word to refer to itself that it could use to refer to another person.

Without this representation, Nico would still be able to learn words. It simply would be unable to apply words to itself that it has learned from human examples. This may not seem so bad, but the deficiency would extend to actions as well. Suppose a human claps and says to such a language-learning robot, "This is called 'clapping.'" The human would be providing a label for human clapping, but not robot clapping if the robot classifies the two activities as separate. Suppose that the human says next, "Now, *clap*." The human would be telling the robot to change a human property, not a robot property, and so the confused robot would simply sit still, or possibly even reach over to put the human's hands together.

There are certainly workarounds in this specific example. Perhaps the robot could be programmed to assume that labels for demonstrated physical activities can always apply to itself, or that non-transitive commands imply that the robot is capable of carrying out the command all by itself. Such workarounds only appear to be delaying the inevitable, however; in general, a word is a word regardless of whether it refers to a human or a robot. It is also not clear what being so conservative will accomplish in the end, since the robot will eventually sense that the laypeople around it refer to it using human-centric terms. A language-learning robot that can represent a concept about itself at all will eventually refer to that concept using the closest human equivalent.

Robots and ethical language

Suppose, then, that there is a robot that can learn language in general, and it learns the word "kill." Moreover, it has been told, "killing is wrong." If the robot is employing a language learning system in which human equivalence is assumed, it will then make the surprising inference that it is wrong to kill robots. This happens because when the robot learns the word *kill*, it has learned it in such a way that applies equally well to humans and robots. Thus, the leap in logic is not in the robot's ethical programming, but in its linguistic understanding.

Of course, it is entirely possible to state every ethical truth in such a way as to avoid linguistic ambiguity: "It is wrong to kill humans," and so on. This is easy, but subtle. The overall effect would be to create a perception that the robots were assuming themselves to be morally equivalent to humans unless told otherwise. This spooky behavior would probably lead to unwarranted assumptions about the machines having developed self-awareness and a growing sense of empower-

ment, when of course the simple reality would be that the robots were merely responding to the same linguistic analogies and metaphors their users had always employed.

It is also worth noting that this issue can crop up for even systems that do not have explicitly ethical programming. A robot able to understand ethical directives is not much different from a robot that responds to commands that happen to be ethical; "do not kill" as a command carries the same human-robot ambiguity as "thou shalt not kill." The difference between an ethical directive and a command is more or less one of time and immediacy.

A second problem, distinct from the linguistic ambiguity of ethical directives, is that robots that can learn to perform speech acts can make ethical claims without actually caring much about the statements' ethical consistency or content. Consider a robot that observes a man being paid. "That's not fair," he might say. "I worked hard – I deserve more than this." The man's employer grudgingly hands over more money. If the robot understands the utility of money, it may approach its own owner and ask for pay. When the owner refuses to pay money to a mere robot, the robot may use the speech act it saw demonstrated: "That's not fair. I worked hard – I deserve more than this." To the robot, the statement is merely a means to an end, but to the world at large, the robot has made a statement of the right to be paid for work.

How bad is this? This depends on one's opinion of the ethical value of robots in general, but there are probably at least some hypothetical robots to which we would not wish to grant rights. The robots that can make the logical leaps described above do not necessarily have to be very complicated beyond their linguistic abilities, and granting them rights when they may be otherwise relatively stupid would probably be a rash thing to do. Meanwhile, though, the robots may continue to speak as if they were human except when explicitly told otherwise. The resulting confusion among the lay populace might be unfortunate. Even the robots' designers may have implemented sufficiently opaque algorithms as to be taken in by the wondrous behavior; it isn't as if artificial intelligence has never made outrageously grandiose claims before.

At this point, the reader may wonder how the benefits mentioned in the previous section could possibly outweigh the drawbacks of this slightly macabre eventuality. But the ambiguity here is an unavoidable consequence of our own ambivalence and inconsistency about the robots' humanity, as reflected in our language. It is a simple fact that most users will apply human-centric language to the robot in some context, and a robot that can learn language will learn that this usage is acceptable.

But things are not as bad as they may seem. The robots' understanding of language will not ultimately change their non-linguistic behavior, except insofar as they misunderstand commands. The robot that demands equal pay from its owner will not go on to assault its owner unless that was

possible in its programming anyway, linguistic confusion or no; in which case, the robot should never have been given that physical ability in the first place. Any robot that can "intentionally" cause a behavior can do so through error instead; it is the responsibility of the robot's designer to ensure that a robot's capacity for harm is absolutely restricted, regardless of the robot's expected behavior. Language-using robots could mutter away, but in the end the designer should either not give a robot the physical capacity to cause harm, or if that is impossible given the application, program as many explicit software safeguards against such behavior as possible.

In fact, if we are concerned about robots causing harm through their utterances alone, ensuring that a robot's ethical safeguards apply to the speech domain as well is probably the most sensible solution to some of these problems. Such safeguards would not prevent the occasional lapse in which the robot appears to value itself as highly as a human, but it could prevent a robot from demanding money from its employer. Speech acts are in the end a special kind of action, and the robot that is not prevented from senselessly asking its owner for money is just as likely to steal the money, if it treats one planning operator as equivalent to another.

Conclusions

There are a few general principles of robotic ethics that this language-learning conundrum lays bare.

The first is that we will not be able to rely on the robots themselves to tell us when they are finally deserving of ethical status. The convenience of a programmed assumption that the robot is analogous to a human is too good to pass up, but the tradeoff is that robots will be making such assumptions long before they are even close to having the other kinds of mental abilities that might make them tempting for ethical considerations. In the end, robot assumptions about ethical status will either be programmed from the beginning or learned from human language, example, and thought. Either way, the decision rests ultimately with humans, whether this is through the algorithms they choose to implement on the robots, their behavior towards the robots, or both. Decisions about whether we must treat robots ethically will not be made any easier once the robots begin to talk about it.

Depending on one's attitude, one may feel a little sorry for the thoughtful, introspective, intelligent robots that may one day be developed after the public has grown accustomed to unintelligent robots that have been crying human-equivalence for years. Such a situation will not occur in my lifetime, and I don't worry too much about it.

The second principle is that ethical considerations are primarily the responsibility of the designer of a robot, and should not be left for the robot to learn at run-time. Ultimately any source of unpredictability is a potential safety concern, and the robot's capacity to cause harm should be physically limited. Nevertheless, the capacity to cause harm

verbally, such as the proverbial crying fire in a crowded theater, underscores the fact that robots that can use speech actions will need safety constraints that go beyond the physical, and deal with the predicted effects of actions. Such principles should be coded into the robot directly rather than conveyed verbally, to avoid ambiguity.

The third principle is that any roboticist has a duty to be very honest with the public about a system's capabilities, and skeptical even in private about the mental abilities of the robot. Unfortunately, the trend toward neural networks, support vector machines, and other machine learning techniques that are not easily analyzable has led many researchers into the trap of not understanding why their own systems work. This can lead to overly optimistic claims. The assumption of human-equivalence is one that I make explicitly, but it is one that many designers of imitation systems might make implicitly – or it may even be "learned" in the sense that somewhere in a neural network jumble this equivalence happens to fall into place. A language-learning robot with such black boxes in its construction that suddenly began to make ethical claims would almost certainly be hailed as a miracle, with nobody really understanding that it was a linguistic assumption driving its behavior.

Robots that can learn language will ultimately reflect human ethics, attitudes, and assumptions in their own speech. It may be worthwhile to decide beforehand what, exactly, we want to tell them.

Acknowledgments

References

Gold, K. and Scassellati, B. (2006a). Deictic pronoun learning and mirror self-identification. In *Proceedings of the 6th International Conference on Epigenetic Robotics (EpiRob-06)*, pages 49–54.

Gold, K. and Scassellati, B. (2006b). Using context and sensory data to learn first and second person pronouns. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction*, pages 110–117, Salt Lake City, Utah.