Dr Alistair Isaac

Should Robots Lie to Us?

This is Forward Thinking, I'm Isabella Melking.

As robots and artificial intelligence play an all the more important role in our lives, should we be worried about how we want them to behave?

Is it ethical for robots to lie to us?

Philosopher of Science Alistair Isaac thinks so! He argues that robots need to be able to lie to us to interact better with humans.

Science communications student Alex Perry talks to Dr Alistair Isaac about his recent paper published in the book Robot Ethics 2.0

Alex:
I am here with Dr Isaac. Dr Isaac has recently published a paper of whether it would be ethical for artificial intelligence to lie to us. Now Dr Isaac I'm sure you're aware of the work of Isaac Asimov who wrote numerous science fiction stories made famous by the film *I Robot* in particular, about the ethics of robots and the idea that robots could be programmed to be ethical and good to humans. But in this context, you have suggested in your paper that it would be a useful feature for robots or artificial intelligence to be able to lie to us. Please could you elaborate on that?

Alistair:
Sure, thank you very much Alex. Yeah, so the paper is all about would sort of situation we'd need to attain if we were to have robots that could interact with us socially that we could treat as colleagues in some sense or interact with smoothly in the workplace. I think that realistically robotics is quite a long way away from that. But it's the sort of thing we imagine happening some day in the relatively near future and that we're actively striving towards in robot development, so it's interesting to think about what we would need for robots to interact smoothly with human beings in the workplace.

And we argue that some ability to deceive, lie, say things that aren't strictly speaking true would be necessary in order for robots to interact socially, because there's all kinds of expressions we use every day when we talk to each other that don't satisfy the strict requirements of truth. So, things like hyperbole for instance. If I say I could eat a horse I don't literally mean I could eat a horse. Now we don't normally think of that as an instance of lying or deception, because we know that it's meant as hyperbole and we're so familiar with these sorts of expressions that we don't bother to step outside and ask "are they literally true?" But if you're thinking about getting an artificial agent to converse, the traditional way to think about how language works in this case is to first analyse what the literal truth of a statement is and then think about what the broader implications are in terms of what you should say next, how you should act. And so, in so far as things like hyperbole, are the sorts of everyday conversational moves that we make with each other,

we're going to need robots to say these sort of things as well, in order for us to be interacting with them just like we interact with other humans.

Now, hyperbole is perhaps an easy case, but in the office workplace it is quite common for people to utter little white lies toward each other, to compliment a colleague perhaps on a new haircut, even if you don't actually like the new haircut. Things go even deeper than that because, there are often situations where one is forced to endorse a superior's decision, even though strictly speaking you don't agree with it. But by doing so you're creating a sort of smoother overall decision-making process and it may even allow you to later on down the line express disagreement on some issue that's more important. So, in the paper, we're considering these ways in which different sorts of technically deceptive behaviour promote an overall positive work environment.

Coming back to Isaac Asimov and I Robot and the sorts of things Asimov talks about, Asimov has been treated as kind of a founding figure, I think, for the ethics of robotics, because he did attempt to lay out what sorts of principles will we need robots to follow in order to ensure that human beings are safe. Things like instilling in them the law not to harm other human beings, instilling in them some sort of law of self-preservation, asking how these two laws might interact.

One of the things we discuss in the paper is that these laws address underlying goals and the robot's surface behaviour is generated in some way by these underlying goals. And the point of the paper is to say look technically deceptive speech - instances where you utter something and you know that it's false and you're trying to convince somebody maybe of it being false - the right way to think about that in the context of the ethics of robotics is not as having the status of being itself an underlying goal but rather as being surface behaviour of some sort that's going to be modulated by underlying goals. And if you're really worried about getting robots to behave ethically, what really matters is getting the underlying goals right, not stipulating some sort of surface behaviour about lying. And in fact, the other way round, it may be that being able to lie is one of those things that's going to be necessary for us to smoothly interact with robots in the workplace.

Alex:
Thank you. Now another aspect that again is of interest to science fiction writers but also, I believe in the philosophy department is the capacity for evolution.

Alistair:
Mmm.

Alex:
That artificial intelligence can evolve just as human behaviour has evolved. Do you see artificial intelligence having the capacity to evolve in the future and could that be a problem for us?

Alistair:
So, the possibility that artificial agents might evolve and especially that they might evolve to be more intelligent than us, more powerful than us, is one of the scenarios that lots of

people interested in robot ethics are worried about. It's a scenario that you see explored in all kinds of science fiction movies. This is what happens in *The Matrix* and *Terminator 2* and it results in terrible catastrophe. Is it a realistic possibility? In some sense the answer is definitely 'yes, because all that's required for evolution to occur is that the conditions of natural selection be there. So, if there are artificial agents which can reproduce themselves somehow and those reproductions can change in slight ways and there's some sort of competition amongst them for survival or something like this then there will be evolution.

Having said that I think we are quite a long way from actual robot evolution or actual evolution of artificial agents. The idea that robots can make other robots is certainly around. Because we use robots of one sort to make robots of other sorts on assembly lines when they're installing electronic equipment in your car for instance, there's some sort of robot that's helping to do that. But nothing that's approaching the level of human intelligence is around yet. And there's no evidence that robots are able to reproduce other robots of greater intelligence than themselves yet. This is something that people have argued about at a theoretical level whether or not it's possible. I think if you look at evolution, natural systems it definitely should be possible because we evolved somehow. But that took a very, very long time. So…

Alex:
Yes, indeed.

Alistair:
I think it's not something we need to be worried about in the short term.

Alex:
OK, well thank you very much for agreeing to be interviewed today. Thank you, bye.

Alistair:
Thank you.

If you want to know more about the Robot Ethics discussed in this podcast follow the links on the Forward Thinking blog at forwardthinking.ppls.ed.ac.uk

Subscribe to our podcast on iTunes for more research, news and views from Philosophy, Psychology and Language Sciences here at the University of Edinburgh.