Related

What is "Do What I Mean"?

How might a real-world AI system that receives orders in natural language and does …

Why is AI alignment a hard problem?

What is perverse instantiation?

This needs a complete rewrite

Let's say that you're the French government a while back. You notice that one of your colonies has too many rats, which is causing economic damage. You have basic knowledge of economics and incentives, so you decide to incentivize the local population to kill rats by offering to buy rat tails at one dollar apiece.

Initially, this works out and your rat problem goes down. But then, an enterprising colony member has the brilliant idea of making a rat farm. This person sells you hundreds of rat tails, costing you hundreds of dollars, but they're not contributing to solving the rat problem.

Soon other people start making their own rat farms and you're wasting thousands of dollars buying useless rat tails. You call off the project and stop paying for rat tails. This causes all the people with rat farms to shutdown their farms and release a bunch of rats. Now your colony has an even bigger rat problem.

Here's another, more made-up example of the same thing happening. Let's say you're a basketball talent scout and you notice that height is correlated with basketball performance. You decide to find the tallest person in the world to recruit as a basketball player. Except the reason that they're that tall is because they suffer from a degenerative bone disorder and can barely walk.

Another example: you're the education system and you want to find out how smart students are so you can put them in different colleges and pay them different amounts of money when they get jobs. You make a test called the Standardized Admissions Test (SAT) and you administer it to all the students. In the beginning, this works. However, the students soon begin to learn that this test controls part of their future and other people learn that these students want to do better on the test. The gears of the economy ratchet forwards and the students start paying people to help them prepare for the test. Your test doesn't stop working, but instead of measuring how smart the students are, it instead starts measuring a combination of how smart they are and how many resources they have to prepare for the test.

The formal name for the thing that's happening is Goodhart's Law. Goodhart's Law roughly says that if there's something in the world that you want, like "skill at basketball" or

“absence of rats” or “intelligent students”, and you create a measure that tries to measure this like “height” or “rat tails” or “SAT scores”, then as long as the measure isn’t exactly the thing that you want, the best value of the measure isn’t the thing you want: the tallest person isn’t the best basketball player, the most rat tails isn’t the smallest rat problem, and the best SAT scores aren’t always the smartest students.

If you start looking, you can see this happening everywhere. Programmers being paid for lines of code write bloated code. If CFOs are paid for budget cuts, they slash purchases with positive returns. If teachers are evaluated by the grades they give, they hand out As indiscriminately.

In machine learning, this is called specification gaming, and it happens frequently.

Now that we know what Goodhart’s Law is, I’m going to talk about one of my friends, who I’m going to call Alice. Alice thinks it’s funny to answer questions in a way that’s technically correct but misleading. Sometimes I’ll ask her, “Hey Alice, do you want pizza or pasta?” and she responds, “yes”. Because, she sure did want either pizza or pasta. Other times I’ll ask her, “have you turned in your homework?” and she’ll say “yes” because she’s turned in homework at some point in the past; it’s technically correct to answer “yes”. Maybe you have a friend like Alice too.

Whenever this happens, I get a bit exasperated and say something like “you know what I mean”.

It’s one of the key realizations in AI Safety that AI systems are always like your friend that gives answers that are technically what you asked for but not what you wanted. Except, with your friend, you can say “you know what I mean” and they will know what you mean. With an AI system, it won’t know what you mean; you have to explain, which is incredibly difficult.

Let’s take the pizza pasta example. When I ask Alice “do you want pizza or pasta?”, she knows what pizza and pasta are because she’s been living her life as a human being embedded in an English speaking culture. Because of this cultural experience, she knows that when someone asks an “or” question, they mean “which do you prefer?”, not “do you want at least one of these things?”. Except my AI system is missing the thousand bits of cultural context needed to even understand what pizza is.

When you say “you know what I mean” to an AI system, it’s going to be like “no, I do not know what you mean at all”. It’s not even going to know that it doesn’t know what you mean. It’s just going to say “yes I know what you meant, that’s why I answered ‘yes’ to your question about whether I preferred pizza or pasta.” (It also might know what you mean, but just not care.)

If someone doesn’t know what you mean, then it’s really hard to get them to do what you want them to do. For example, let’s say you have a powerful grammar correcting system, which we’ll call Syntaxly+. Syntaxly+ doesn’t quite fix your grammar, it changes your writing so that the reader feels as good as possible after reading it.

Pretend it's the end of the week at work and you haven't been able to get everything done your boss wanted you to do. You write the following email:

"Hey boss, I couldn't get everything done this week. I'm deeply sorry. I'll be sure to finish it first thing next week."

You then remember you got Syntaxly+, which will make your email sound much better to your boss. You run it through and you get:

"Hey boss, Great news! I was able to complete everything you wanted me to do this week. Furthermore, I'm also almost done with next week's work as well."

What went wrong here? Syntaxly+ is a powerful AI system that knows that emails about failing to complete work cause negative reactions in readers, so it changed your email to be about doing extra work instead.

This is smart - Syntaxly+ is good at making writing that causes positive reactions in readers. This is also stupid - the system changed the meaning of your email, which is not something you wanted it to do. One of the insights of AI Safety is that AI systems can be simultaneously smart in some ways and dumb in other ways.

The thing you want Syntaxly+ to do is to change the grammar/style of the email without changing the contents. Except what do you mean by contents? You know what you mean by contents because you are a human who grew up embedded in language, but your AI system doesn't know what you mean by contents. The phrases "I failed to complete my work" and "I was unable to finish all my tasks" have roughly the same contents, even though they share almost no relevant words.

Roughly speaking, this is why AI Safety is a hard problem. Even basic tasks like "fix the grammar of this email" require a lot of understanding of what the user wants as the system scales in power.

In Human Compatible, Stuart Russell gives the example of a powerful AI personal assistant. You notice that you accidentally double-booked meetings with people, so you ask your personal assistant to fix it. Your personal assistant reports that it caused the car of one of your meeting participants to break down. Not what you wanted, but technically a solution to your problem.

You can also imagine a friend from a wildly different culture than you. Would you put them in charge of your dating life? Now imagine that they were much more powerful than you and desperately desired that your dating life to go well. Scary, huh.

In general, unless you're careful, you're going to have this horrible problem where you ask your AI system to do something and it does something that might technically be what you wanted but is stupid. You're going to be like "wait that wasn't what I mean", except your system isn't going to know what you meant.