How to Build a Robot That Wants to Change the World

The computer scientist Christoph Salge is trying to circumvent the need for rules that guide robots' behavior. His strategy: Give them a goal of making us more powerful.

By John Pavlus

Sasha Maslov for Quanta Magazine

Christoph Salge at New York University's Game Innovation Lab.

Isaac Asimov's famous Three Laws of Robotics — constraints on the behavior of androids and

automatons meant to ensure the safety of humans — were also famously incomplete. The laws, which first appeared in his 1942 short story "Runaround" and again in classic works like *I, Robot*, sound airtight at first:

> 1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
> 2. A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.
> 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Of course, hidden conflicts and loopholes abound (which was Asimov's point). In our current age of advanced machine-learning software and autonomous robotics, defining and implementing an airtight set of ethics for artificial intelligence has become a pressing concern for organizations like the Machine Intelligence Research Institute and OpenAI.

Christoph Salge, a computer scientist currently at New York University, is taking a different approach. Instead of pursuing top-down philosophical definitions of how artificial agents should or shouldn't behave, Salge and his colleague Daniel Polani are investigating a bottom-up path, or "what a robot should do in the first place," as they write in their recent paper, "Empowerment as Replacement for the Three Laws of Robotics." Empowerment, a concept inspired in part by cybernetics and psychology, describes an agent's intrinsic motivation to both persist within and operate upon its environment. "Like an organism, it wants to survive. It wants to be able to affect the world," Salge explained. A Roomba programmed to seek its charging station when its batteries are getting low could be said to have an extremely rudimentary form of empowerment: To continue acting on the world, it must take action to preserve its own survival by maintaining a charge.

Empowerment might sound like a recipe for producing the very outcome that safe-AI thinkers like Nick Bostrom fear: powerful autonomous systems concerned only with maximizing their own interests and running amok as a result. But Salge, who has studied human-machine social interactions, wondered what might happen if an empowered agent "also looked out for the empowerment of another. You don't just want your robot to stay operational — you also want it to maintain that for the human partner."

Salge and Polani realized that information theory offers a way to translate this mutual empowerment into a mathematical framework that a non-philosophizing artificial agent could put into action. "One of the shortcomings of the Three Laws of Robotics is that they are language-based, and language has a high degree of ambiguity," Salge said. "We're trying to find something that is actually operationizable."

Quanta spoke with Salge about information theory, nihilist AI and the canine model of human-robot interaction. An edited and condensed version of the conversation follows.

Some technologists believe that AI is a major, even existential threat. Does the prospect of runaway AI worry you?

I'm a bit on the fence. I mean, I do think there are currently genuine concerns with robots and the growing influence of AI. But I think in the short term we're probably more concerned about maybe job replacement, decision making, possibly a loss of democracy, a loss of privacy. I'm unsure how likely it is that this kind of runaway AI will happen anytime soon. But even an AI controlling your health care system or what treatment options you're getting — we should start to be concerned

about the kind of ethical questions that arise from this.

How does the concept of empowerment help us deal with these issues?

I think that the idea of empowerment does fill a niche. It keeps an agent from letting a human die, but once you've satisfied this very basic bottom line, it still has a continued drive to create additional possibilities and allow the human to express themselves more and have more influence on the world. In one of Asimov's books, I think the robots just end up putting all the humans in some kind of safe containers. That would be undesirable. Whereas having our abilities to affect the world continuously enhanced seems to be a much more interesting end goal to reach.

You tested your ideas on virtual agents in a video game environment. What happened?

An agent motivated by its own empowerment would jump out of the way of a projectile, or keep from falling into a hole, or avoid any number of situations that would result in its losing mobility, dying or being damaged in a way that would reduce its operationality. It just keeps itself running.

When it was paired with a human player that it was supposed to empower as well as itself, we observed that the virtual robot would keep a certain distance so as not to block the human's movement. It doesn't block you in; it doesn't stand in a doorway that's then impossible for you to pass through. We basically saw that this effect keeps the companion sticking close to you so it can help you out. It led to behavior where it could take the lead or follow.

For example, we also created a scenario where we had a laser barrier that would be harmful for the human, but not harmful for the robot. If the human in this game gets closer to the laser, suddenly there is more and more of an empowerment-driven incentive for the robot to block the laser. The incentive gets stronger when the human stands right next to it, implying, "I want to cross this now." And the robot would actually block the laser by standing in front of it.

Did the agents engage in any unintended behavior, like the kind that emerges from the three laws in Asimov's fiction?

We initially got good behavior. For example, the virtual robot takes out enemies that are trying to kill you. Once in a while it might jump in front of a bullet for you, if this is the only way to save you. But one thing that was a bit surprising to us, at the beginning, was that it was also very afraid of you.

The reason for this has to do with its "local forward" model: Basically, it looks at how certain action sequences two or three steps into the future affect the world, for both you and itself. So as a first, easy step, we programmed this model to assume that the player would act randomly. But in practice, that meant that the agent was essentially acting under the assumption that the human player is kind of a psychopath, and so at any point in time that human could decide to, for example, fire at the agent. So the agent would always be very, very careful to be in positions where the human couldn't kill it.

We had to fix this, so we modeled something we call a trust assumption. Basically, the companion agent acts under the assumption that the human will only choose those actions that will not remove the agent's own empowerment — which is probably a more natural model for a companion anyway.

The other thing we noticed in the game was that, if you had, say, 10 health points, the companion wasn't really concerned with you losing the first eight or nine of these — and would even shoot you once in a while just for laughs. There, again, we realized that there's a disconnect between the world we live in and the model in a computer game. Once we modeled a limitation of ability resulting from health loss, this problem went away. But it also could have been dealt with by designing the local-forward model in a way that makes it able to look further into the future than just a few steps. If the agent were able to look really far into the future, it would see that having more health points might be helpful for the things to come.

Whereas if the loss of spare health points doesn't make a difference to my empowerment right now …

The agent basically goes, "Oh, I could not shoot him, or I could shoot him. No difference." And sometimes it shoots you. Which of course is a problem. I do not condone the random shooting of players. We've added a fix so the virtual robot cares a bit more about your empowerment than about its own.

How do you make these concepts precise?

If you think about agents as control systems, you can think in terms of information: Stuff happens in the world, and this somehow affects you. We're not just talking about information in terms of things you perceive, but as any kind of influence — it could be matter, anything flowing back and forth between the world and you. It might be the temperature affecting you, or nutrients entering your body. Any kind of thing that permeates this boundary between the world and the agent carries information in. And in the same way, the agent can affect the outside world in numerous ways, which also outputs information.

You can look at this flow as a channel capacity, which is a concept from information theory. You have high empowerment if you have different actions you can take that will lead to different results. If any of these capabilities become worse, then your empowerment goes down — because the loss of capability corresponds with a quantifiable reduction in this channel capacity between you and the environment. This is the core idea.

How much does the agent need to know for empowerment to work?

Empowerment has the advantage that it can be applied even if your knowledge isn't complete. The agent does need a model of how its actions are going to affect the world, but it doesn't need a complete understanding of the world and all its intricacies. In contrast to some approaches that try to model everything in the world as best they can and then try to figure out what their actions actually mean, here you only need to figure out how your actions affect your own perception. You don't have to figure out where everything is; you can have an agent that explores the world. It does things and tries to figure out how its actions affect the world. As this model grows, the agent also gets better at figuring out how empowered it is.

You've tested this in virtual environments. Why not the real world?

The main obstacle to scaling this model up, and why we're not putting this on any real robot yet, is that it's hard to compute the channel capacity of an agent and a human far forward in time in a rich environment like the real world. There are a lot of initiatives under way to make this more efficient. I'm optimistic, but currently it is a computational concern. That's why we applied the framework to a computer game companion, which of course is a much more simplistic form, making the

computational issues easier to solve.

It sounds as if empowerment, ideally, would make our machines act like really powerful service dogs.

I actually know some roboticists who are deliberately modeling companion behavior after dogs. I mean, having robots treat us like our dogs treat us is probably a future we can all live with.