How can humans keep the upper hand on artificial intelligence?

December 1 2017, by Anne-Muriel Brouet

From right to left: Rachid Guerraoui, Alexandre Maurer, El Mahdi El Mhamdi. Credit: ©Alain Herzog/EPFL

EPFL researchers have shown how human operators can maintain control over a system comprising several agents that are guided by

artificial intelligence.

In artificial intelligence (AI), machines carry out specific actions, observe the outcome, adapt their behavior accordingly, observe the new outcome, adapt their behavior once again, and so on, learning from this iterative process. But could this process spin out of control? Possibly. "AI will always seek to avoid human intervention and create a situation where it can't be stopped," says Rachid Guerraoui, a professor at EPFL's Distributed Programming Laboratory and co-author of the EPFL study. That means AI engineers must prevent machines from eventually learning how to circumvent human commands. EPFL researchers studying this problem have discovered a way for human operators to keep control of a group of AI robots; they will present their findings on Monday, 4 December, at the Neural Information Processing Systems (NIPS) conference in California. Their work makes a major contribution to the development of autonomous vehicles and drones, for example, so that they will be able to operate safely in numbers.

One machine-learning method used in AI is reinforcement learning, where agents are rewarded for performing certain actions – a technique borrowed from behavioral psychology. Applying this technique to AI, engineers use a points system where machines earn points by carrying out the right actions. For instance, a robot may earn one point for correctly stacking a set of boxes and another point for retrieving a box from outside. But if, on a rainy day for example, a human operator interrupts the robot as it heads outside to collect a box, the robot will learn that it is better off staying indoors, stacking boxes and earning as many points as possible. "The challenge isn't to stop the robot, but rather to program it so that the interruption doesn't change its learning process – and doesn't induce it to optimize its behavior in such a way as to avoid being stopped," says Guerraoui.

From a single machine to an entire AI network

In 2016, researchers from Google DeepMind and the Future of Humanity Institute at Oxford University developed a learning protocol that prevents machines from learning from interruptions and thereby becoming uncontrollable. For instance, in the example above, the robot's reward – the number of points it earns – would be weighted by the chance of rain, giving the robot a greater incentive to retrieve boxes outside. "Here the solution is fairly simple because we are dealing with just one robot," says Guerraoui.

However, AI is increasingly being used in applications involving dozens of machines, such as self-driving cars on the road or drones in the air. "That makes things a lot more complicated, because the machines start learning from each other – especially in the case of interruptions. They learn not only from how they are interrupted individually, but also from how the others are interrupted," says Alexandre Maurer, one of the study's authors.

Hadrien Hendrikx, another researcher involved in the study, gives the example of two self-driving cars following each other on a narrow road where they can't pass each other. They must reach their destination as quickly as possible – without breaking any traffic laws – and humans in the cars can take over control at any time. If the human in the first car brakes often, the second car will adapt its behavior each time and eventually get confused as to when to brake, possibly staying too close to the first car or driving too slowly.

Giving humans the last word

This complexity is what the EPFL researchers aim to resolve through "safe interruptibility." Their breakthrough method lets humans interrupt AI learning processes when necessary – while making sure that the interruptions don't change the way the machines learn. "Simply put, we

add 'forgetting' mechanisms to the learning algorithms that essentially delete bits of a machine's memory. It's kind of like the flash device in Men in Black," says El Mahdi El Mhamdi, another author of the study. In other words, the researchers altered the machines' learning and reward system so that it's not affected by interruptions. It's like if a parent punishes one child, that doesn't affect the learning processes of the other children in the family.

"We worked on existing algorithms and showed that safe interruptibility can work no matter how complicated the AI system is, the number of robots involved, or the type of interruption. We could use it with the Terminator and still have the same results," says Maurer.

Today, autonomous machines that use reinforcement learning are not common. "This system works really well when the consequences of making mistakes are minor," says El Mhamdi. "In full autonomy and without human supervision, it couldn't be used in the self-driving shuttle buses in Sion, for instance, for safety reasons. However, we could simulate the shuttle buses and the city of Sion and run an AI algorithm that awards and subtracts points as the shuttle-bus system learns. That's the kind of simulation that's being done at Tesla, for example. Once the system has undergone enough of this learning, we could install the pre-trained algorithm in a self-driving car with a low exploration rate, as this would allow for more widespread use." And, of course, while making sure humans still have the last word.

More information: Dynamic Safe Interruptibility for Decentralized Multi-Agent Reinforcement Learning, arxiv.org/pdf/1704.02882.pdf

Provided by Ecole Polytechnique Federale de Lausanne