Daniel E. Geer, Jr.
In-Q-Tel

Unknowable Unknowns

Kurt Gödel proved that there are problems for which it is impossible to construct an algorithm that always leads to a correct yes-or-no answer; those problems are undecidable. Alan Turing proved that the halting problem is undecidable in Turing machines. Alfred Tarski proved that truth in the standard model of a system cannot be defined within that system. Olav Lysne proved that it is not possible to verify electronic equipment procured from untrusted vendors, and that a vendor cannot build a system that supports verification by untrusting customers. Ben-David et al. proved that scenarios exist where learnability can neither be proved nor refuted. Finally, Donald Rumsfeld made commonplace the phrase *unknown unknowns*.

And so we come to artificial intelligence (AI), which is to say self-modifying algorithms, which is to say machine learning. Readers of *IEEE Security & Privacy* are well aware of the interrogatability problem "Monsieur Algorithme, why did you make this decision?" an acute concern in multiple subject matter areas, of which cybersecurity is assuredly one. Most *IEEE Security & Privacy* readers agree that all security tools are dual-use, freighting "Why did you make this decision?" with significantly more than mere curiosity or a search for optimality.

There are some who say that a self-modifying algorithm, if purposefully and skillfully constructed, can tell the "why" of its decisions, tell that why in a form we humans can appreciate, and then, perhaps, nod in knowing acceptance. One hopes that this will soon be true, but as of now, it is not.

In other words, and for the time being, black-box interrogation of AI models—similar in spirit to a statistician's sensitivity testing—has the potential to become the default method of assessing an AI model's behaviors. This default will last at least so long as no one has success in understanding a priori how a model works. It is merely stop-gap, relative to Rumsfeld's unknown unknowns: the probability of not asking enough of the right type of questions to characterize a data-driven model is as great as the probability that the model was trained on incomplete or biased data.

It is logical to presume that as AI models increase in complexity, they become more opaque. This parallels a problem we in cybersecurity know only too well: that of trying to understand the attack surface of growing and/or dynamic software installations. Unprovability thus becomes acute, including in the case of cybersecurity, where the mutation rate for offense and defense alike mean not just learning but unlearning.

In some areas other than cybersecurity, handing off the keys to AI models offers immediate, iterative improvement in tailored operations, efficiency, and safety. In the arms race that is cybersecurity, using adaptive algorithms to thwart other adaptive algorithms is so attractive as to seem necessary, and so necessary as to seem attractive.

It is logical to presume that as AI models increase in complexity, they become more opaque.

The financial services industry has already demonstrated some apparent truths worth considering, the principal of which is that we (humans) can build systems more complex than we can manage, complete with behaviors that we cannot predict. Perhaps the question is whether self-modification is, or can be made to be, a safe enough technology to implement, and if so, how does this decision vary by the realm of application?

Digital Object Identifier 10.1109/MSEC.2019.2898636
Date of publication: 2 April 2019

continued on p. 79

 Copublished by the IEEE Computer and Reliability Societies

 continued from p. 80

In human society, it is natural for the occasional interrogator who asks "Why did you do that?" to demand an action reversal based on the answer to the question. In the digital policy world, Article 15, Section 1(h) of the European Union's General Data Protection Regulation reads "The data subject shall have the right to obtain from the controller (…) access to personal data and the following information: 1(h), the existence of automated decision making [and] meaningful information about the logic involved as well as the significance and the envisaged consequences of such processing for the data subject." Cybersecurity decisions will certainly encounter Article 15's requirement, and for cybersecurity services that only know what to interdict by being trained on "normal day" data, there is no real answer to Section 1(h)'s requirement as to whether there was hidden malignancy in the training data—i.e., that is an unknowable unknown.

The author suggests that an exclusive embrace of machine learning for cybersecurity is a Faustian bargain—but it's a free country. ■

Daniel E. Geer, Jr. is the chief information security officer of In-Q-Tel. Contact him at dan@geer.org.

Digital Object Identifier 10.1109/MSEC.2019.2903679