Elisa Bertino
Purdue University

Attacks on Artificial Intelligence

There is, today, a lot of hype around the issue of attacks on artificial intelligence (AI). There are huge numbers of research papers, white papers and reports about those attacks and potential defenses—which is reasonable given that AI techniques are today pervasive in all applications we may think of. As we increasingly rely on AI techniques for critical decisions, forecasts and analyses, concerns about whether these AI techniques can be attacked are certainly legitimate.

Well-known attacks include input attacks and poisoning attacks. In an input attack, the attacker manipulates the data that is fed to the AI algorithm to manipulate the output of the algorithm. A lot of work concerning this type of attack has been done for neural networks. In a poisoning attack, the attacker tampers with the process by which the AI algorithm is trained. For example, the attacker can corrupt the data used during training, so that the algorithm will misclassify certain instances.

While I believe that focusing on specific attacks and defenses is important, I also believe that we need to look at a broader picture. Consider the case of input attacks. When you look at those attacks, an example often given is the one where you have an autonomous vehicle running a neural network to recognize stop signs and an attacker that has partially covered a stop sign so that the neural network is not able to recognize it. My objection to such an example is that autonomous vehicles will most likely have maps on-board with stop signs marked in these maps; stop signs in the future may also emit sounds, and there will be vehicle-to-vehicle communications and information transmission among vehicles. In other words, vehicles will have multiple information sources that can be compared and correlated so that the correct and safe decision is taken. The idea of comparing/fusing multiple inputs from different sources has been around for a very long time—see the many techniques proposed in the data fusion area, which started in the early 1980s. My point here is that we should look at the problem of AI security from a system point of view by which you want to make your system secure and reliable by including different AI techniques and models, and using data from independent sources. Such an approach would not only enhance security, but also the timeliness and coverage of decisions, predictions, etc. In other words, our ultimate goal should not be just the protection of the AI itself, rather it should be to make accurate decisions, forecasts and analyses and to achieve this goal we need to think in terms of systems security.

Consider now the case of poisoning attacks. A defense against those attacks is to deploy well-known data security practices and data provenance techniques. Data which is of poor quality or altogether malicious affects not only AI, but most applications we may think of. For example, malicious data injection attacks have been shown for control systems that negatively affect decisions taken by these systems, especially now that these systems are becoming interconnected with other systems. So, it would seem to me that if we are able to secure data and securely record and manage its provenance, we would be able to protect AI against data poisoning attacks.

However, we still want to do the best we can for securing AI, very much like we try to do for software. We thus need to came up with suitable AI assurance processes, which perhaps can be based on extensions to software assurance processes. AI assurance processes would, of course, need to include assurance about the data used for training and testing and also assurance about the software implementing the AI

Digital Object Identifier 10.1109/MSEC.2020.3037619
Date of current version: 25 January 2021

continued on p. 103

1540-7993/21©2021IEEE

Last Word

continued from p. 104

techniques. Undoubtedly, there will be additional challenges, for example, in the case of AI techniques, such as reinforcement learning, by which AI algorithms dynamically modify their behavior based on experience with environments. Also, different AI assurance processes will be required for different application domains, such as for safety critical domains—very much as it done today for software that has to be deployed in safety-critical applications.

To conclude, in my view, AI security is essential but has to be addressed with a system perspective in mind and by developing suitable assurance processes, and last but not least by making sure that data is secure and trustworthy. In this respect, data transparency is a critical building block. Just focusing on attacks that flip some bits in images or tamper training would not be sufficient. Certainly, there is also the problem of AI privacy, but this discussion will be for another time. ■

Elisa Bertino is a professor with Purdue University, West Lafayette, Indiana, 47907, USA. Contact her at bertino@purdue.edu.

Errata

In the November/December 2020 issue of *IEEE Security & Privacy*, there was an error in a biography in Padilha et al.[1] The bio of Gabriel Bertocco should read as follows:

Gabriel Bertocco is currently pursuing a Ph.D. in computer science, with a focus on digital forensics, at the University of Campinas, Brazil, where he received a B.Sc. in computing engineering in 2019. His research interests include machine learning, computer vision, and digital forensics. Contact him at gabriel.bertocco@ic.unicamp.br.

Digital Object Identifier 10.1109/MSEC.2020.3040154
Date of current version: 25 January 2021

Reference

1. R. Padilha, C. M. Rodrigues, F. Andaló, G. Bertocco, Z. Dias, and A. Rocha, “Forensic event analysis: From seemingly unrelated data to understanding,” *IEEE Security Privacy*, vol. 18, no. 6, pp. 23–31, Nov./Dec. 2020. doi: 10.1109/MSEC.2020.3000446.