Researchers suggest medical AI systems could be vulnerable to adversarial attacks

22 March 2019, by Bob Yirka

Credit: CC0 Public Domain

A small team of medical researchers from Harvard University and MIT has published a Policy Forum piece in the journal *Science* suggesting that future medical AI systems could be vulnerable to adversarial attacks. They point out that prior research has shown that virtually all AI systems are vulnerable in some way to such attacks.

An adversarial attack in the field of machine learning is an attempt through malicious input to fool the model upon which such a system is built. In practice, this means feeding an AI system some sort of information that forces it to return incorrect results. The researchers suggest such an attack could be aimed at detection systems like those programmed to find cancer by analyzing scans. They even showed how an adversarial attack would work by feeding a system a certain noise pattern that triggered confusion, resulting in incorrect results.

But this is not the kind of adversarial attack the researchers are really worried about. What most concerns them are the AI systems that have been developed and are in use already that are involved in processing claims and billing—the possibility that hospitals or even doctors could use such systems to alter information on forms to get paid more by insurance companies or Medicaid for carrying out tests, for example, by changing a code to make a simple X-ray look like an MRI test. Feeding an AI system the right piece of information at the right time could make it do just that. There also exists the possibility that a hospital could teach its AI system to find the best ways to scam insurance companies or the government, making it almost impossible to detect.

The researchers suggest that a new approach to policy-making is required—one in which people from a wide variety of fields, including law, computer science and medicine, address the problem before it becomes prevalent. Such groups could, perhaps, find ways to prevent it from happening, or at least detect it if it does.

More information: Samuel G. Finlayson et al. Adversarial attacks on medical machine learning, *Science* (2019). DOI: 10.1126/science.aaw4399