SOPHOS
Cybersecurity made simple.

# "Known Unknowns": Overcoming Catastrophic Failure Modes in Artificial Intelligence

By Dr. Ethan M. Rudd, Sophos Data Scientist
Twitter: @EthanMRudd

"It ain't what you know that gets you in trouble, it's what you know for sure that just ain't so."

A Sophos Whitepaper May 2018

In cybersecurity, machine learning (machine learning) is often portrayed as a godsend that will deliver us to a heavenly place free of signatures. The theory is that machine learning models, given enough training data, can learn to generalize concepts of "malicious" and benign." Inasmuch as this holds, machine learning-based antimalware is proactive because, based off some latent notion of "malicious" behavior that it has learned, it can correctly categorize heretofore unseen malware as such.

Signatures, by contrast, are inherently reactive because they are built from malware in the wild after a number of infections have already occurred. While there is compelling evidence of Machine learning's ability to generalize novel malware types and tremendous evidence to suggest that machine learning is a useful tool in the cybersecurity arsenal, there is equally compelling evidence to suggest that it is no silver bullet and that signatures will also be around for quite some time.

I am not just talking about cherry-picking failure cases and making vigorous hand-wavy statements about a lack of good data, labeling, and sampling. There is a far more fundamental problem with the formulation of machine learning models, which could be construed as a grave design flaw which separates "artificial intelligence" from actual intelligence.

Machine learning models aim to operate in the future, making decisions about novel inputs, yet they are trained with no fundamental notion of uncertainty. They are "taught" during training to categorize certain samples under the assumption that the sample data in question represents an omniscient view of the universe, and that observations will not change. But that's not the way intelligence works in the real world. People, no matter how smart, are rarely certain, and rightly so, because they usually lack complete information.

When the unexpected happens it can be extremely disruptive, because normal paradigms no longer apply. In his bestseller, "The Black Swan," Nassim Taleb contends that, for this reason, a great deal of human history has been shaped by unexpected rare events, whereby models and paradigms that work well for previously seen events fail to extrapolate forward. Consider two unexpected events in recent history that completely changed the state of the world:

- The 9/11 terrorist attacks are still shaping travel policies and US involvement in the Middle East to this day.
- The 2007/2008 housing bust led to an unprecedented global financial crisis.

While not of quite the same impact on the global scale, zero-day cyber attacks have decimated multi-billion dollar companies. In short, unexpected rare occurrences often lead to extremely high-stakes situations, so any intelligent system should at least be able to recognize such occurrences. As Donald Rumsfeld put it, in a seminal press conference, "there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tend to be the difficult ones." Thus, it is in our best interest to know when we do not know.

Unfortunately, most machine learning models, including deep learning models, are designed with no notion of “known unknown”. Another way of saying this is that they are closed-set models, attempting to operate in a world that is inherently open-set. Humans, by contrast, encounter unfamiliar concepts on a daily basis and it is our ability to distinguish the unfamiliar from the familiar that allows us to learn.

Addressing the open set problem is beyond the capability of most machine learning algorithms because their optimization objectives only aim to minimize misclassification of training examples. Viewing this from a risk management paradigm, this is often described as **empirical risk minimization** (Ethan M. Rudd, 2017), where the empirical risk is proportional to the number of misclassified training examples.

In a world governed by fictitious closed set assumptions, empirical risk minimization works well, but in the real open set world, there is more to it. Consider a classifier trained to separate three different classes of data under an empirical risk minimization objective (see Fig. 1). It chops the sample space using piecewise linear boundaries into three subspaces of infinite span. This separates classes of data quite well, as long as points from these three classes of data are all it sees. Moreover, a common practice in this regime is to assess probability of class membership or classifier confidence proportionally to the distance from the decision boundary within the member class.

When data from a novel class occurs, however, it will be classified as belonging to one of the known classes, and when it lies especially far from any of the training samples (see. Fig. 1 left,) it will be classified as one of the known classes with high probability. Thus, the classifier will not only be wrong – it will be very wrong, yet very confident in its decision.

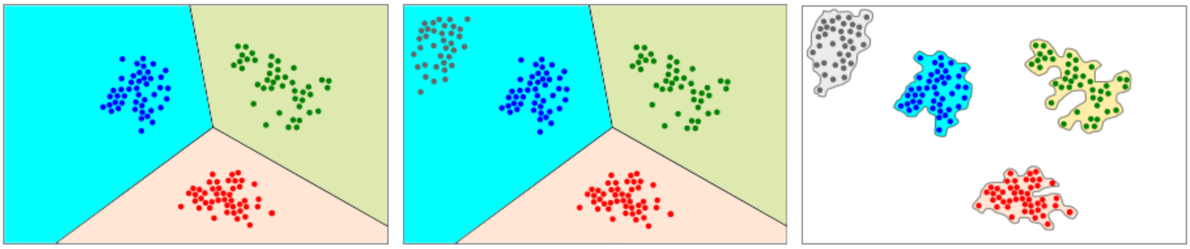

Fig. 1. Left: Three classes of data, blue squares, green triangles, and red diamonds, are separated by a linear classifier. Shading of regions indicates the classifier’s decision, which perfectly separates the data. Middle: when the classifier sees novel data from a different class, it will incorrectly classify it as blue squares. When distance from the decision boundary is used as a confidence/calibration measure, it will classify the novel data as the “blue square” class with even higher probability than the legitimate blue squares. Right: an open set classifier trained to minimize both empirical risk with open space risk would bound class confidence to only label regions with sufficient support from the training set as belonging to a given class.

While the toy example presented in Fig. 1 is simple, low-dimensional, and assumes a linear classifier, intricate, high-dimensional, nonlinear classifiers, including deep neural networks, are also susceptible to the open set problem because there is no term in the optimization process that penalizes ascribing unlabeled space to a particular known class. Moreover, since the density of sample points to unlabeled space can decrease exponentially with dimensionality, there is a lot more “unknown” space to mis-ascribe as known. Summarily, empirical risk minimization is not enough, and, while different modes of failure will occur from adding a more intricate classifier, there is no reason to assume that doing so will mitigate the open set problem.

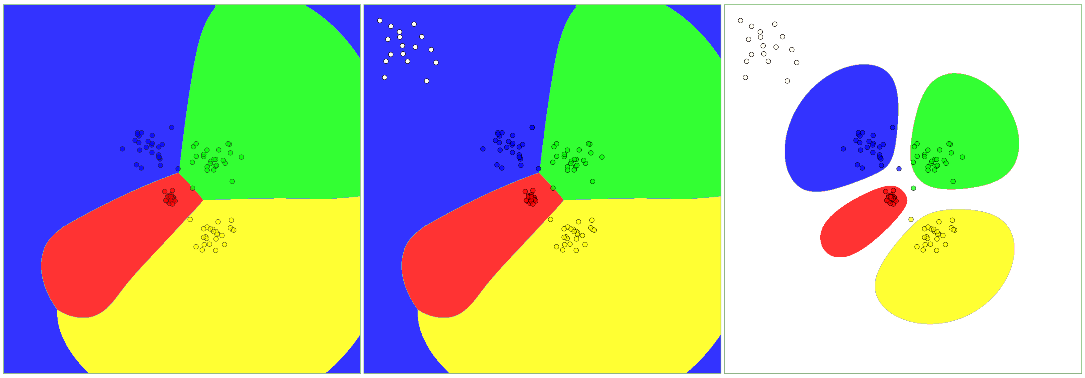

Fig. 2. The open set problem is not a function of a classifier's representational power. Left and Middle: a plausible representation learned by a deep neural network that minimizes empirical risk yet does not factor in open space risk. Right: A balance of empirical risk with open space risk would prevent the classifier from labelling unknown space.

While there are far more hurdles to designing truly intelligent AI systems than the open set problem alone, is it possible to formulate classifiers that avoid making such stupid mistakes? The answer is yes! This can be accomplished by adopting a more realistic risk management framework – one that balances both **empirical risk**, the risk of misclassifying a sample, with **open space risk** – the risk of labeling unknown space (Walter J. Scheirer A. R., 2013).

The balance thereof becomes an **open set risk minimization** problem, solutions to which learn both to classify data in regions of bounded support from training samples, but also learn when there is no basis for making a decision. Several open set decision machines have been pioneered, including the one-versus-set machine (Walter J. Scheirer A. R., 2013), the W-SVM (Walter J. Scheirer L. P., 2014), and the Extreme Value Machine (EVM) (Ethan M. Rudd, 2017).

These operate by explicitly bounding open space risk estimated using cross-class validation, leaving one class out at a time during training and treating these samples as “unknown”, or modeling probability of sample inclusion with respect to a given class.

However, interest in the open set classification regime has only recently become widespread. This is perhaps because, while the problem is simple, the nuances of the formulation are somewhat outside the realm of traditional machine learning. Moreover, unlike “adversarial samples/modeling” and generalization," which have become fads in the machine learning community to the point where much of the underlying research is pedantic, unsubstantiated, and grossly inapplicable, such a following has not yet occurred with open set recognition.

There have also been similar ad-hoc approaches that theoretically bound open-space risk – often doing probability density estimation on the data, then exercising a reject option on the final classification. However, such techniques, when performed, are usually afterthoughts, and the bounds on “open space” are loose and ill-grounded, partly because “probability of class inclusion” and class sample probability density are not the same thing. Moreover, two different models have been fit on separate loss functions and not jointly optimized.

However, there are many immediate use cases for open set classifiers in the security community. For example, many false positives and false negatives that lie in previously unseen regions of hypothesis space could potentially be mitigated. Forcing a classifier to make a decision about a completely foreign type of code can, for example, result in strange behavior. For instance, in Fall 2017 many state-of-the-art vendors recognized a compiled “hello world” binary as malicious, likely because the program was **far too simple** to be similar to any other in the training set. Likewise, zero-day attacks can be easily missed, precisely because their code/behavior patterns differ so substantially from those previously seen.

An open set classifier could potentially detect these and flag them for inspection, creation of signatures, and eventual update to the classifier. More generally, an open set classifier can serve as a tool to detect misclassifications of "known" data, on the one hand suggesting an ill-trained classifier vs. the need to acquire more training data, as well as **what type** of training data to acquire.

There are also similar problems related to open set. For example, when viewing things from an open set paradigm, there is evidence to substantiate that, particularly for deep neural networks, "fooling samples" (Nguyen, 2015) occur due to classification decisions on unsupported space. These are easily recognized and can be labeled as "unknown" by an open set classifier. "Adversarial samples" (Goodfellow, 2015), by contrast, which appear similar to the original inputs in hypothesis space.

Can open set recognition be extended to turn these samples into "known unknowns"? Maybe, maybe not, but it is an interesting topic for future research.

# References

Ethan M. Rudd, L. P. (2017). The Extreme Value Machine. IEEE Transactions on Pattern Recognition and Machine Intelligence. Available: https://arxiv.org/pdf/1506.06112.pdf

Goodfellow, I. J. (2015). Explaining and Harnessing Adversarial Examples. International Conference on Learning Representations. Available: https://arxiv.org/pdf/1412.6572.pdf

Nguyen, A. J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. IEEE Computer Vision and Pattern Recognition. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/app/1A_047.pdf

Walter J. Scheirer, A. R. (2013). Towards Open Set Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. Available: https://www.wjscheirer.com/papers/wjs_tpami2013_openset.pdf

Walter J. Scheirer, L. P. (2014). Probability Models for Open Set Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. Available: https://www.wjscheirer.com/papers/wjs_tpami2014_probability.pdf

United Kingdom and Worldwide Sales
Tel: +44 (0)8447 671131
Email: sales@sophos.com

North American Sales
Toll Free: 1-866-866-2802
Email: nasales@sophos.com

Australia and New Zealand Sales
Tel: +61 2 9409 9100
Email: sales@sophos.com.au

Asia Sales
Tel: +65 62244168
Email: salesasia@sophos.com