From: Ari Gesher <[redacted]
To: Jeffrey Epstein <jeevacation@gmail.com>
Subject: Re: MDF
Date: Thu, 24 Oct 2013 00:37:50 +0000

The adaptive adversaries in the immune system context are parasites. I wrote a piece using biological parasites as a strong analogy for the problems cyber security (which also goes into the need to use side-channels to find and stop adaptive adversaries). The immune system does a horrible job of dealing with parasites (definitionally) and can be seen to be just like standard pattern-matching/data mining/statistical approaches as a defense mechanism - they're really good at stopping what they've seen before and pretty bad at stopping novel attacks.

The self-recognition piece is interesting, though.

On Oct 23, 2013, at 5:26 PM, Jeffrey Epstein <jeevacation@gmail.com> wrote:

I think the immune system might provide useful insights . Recognizing self , attacking everything that is not self

On Wednesday, October 23, 2013, Ari Gesher wrote:

On Oct 23, 2013, at 8:09 AM, Joscha Bach <[redacted]> wrote:

That being said, AGI will have trouble succeeding because it is following the scruffy tradition. Perhaps the main failing of this tradition is its refusal to define objective (and preferably quantitative) measures of success.

The question of good benchmark tasks is haunting AI since its inception. Usually, when we identify a task that requires intelligence in humans (playing chess or soccer or Jeopardy, driving a car etc.) we end up with a kind of very smart (chess-playing, car-driving) toaster. That being said, AI was always very fruitful in the sense that it arguably was the most productive and useful field of computer science, even if it fell short of its lofty goals. Without a commitment to understanding intelligence and mind itself, AI as a discipline may be doomed, because it will lose the cohesion and direction of a common goal.

So now this gets interesting and starts to point us towards both MDF and the study of deception.

The smart toasters emerge because they're being designed to solve well-bounded problems (like playing chess). There is no deception in chess (I would put feints in a different category), no hidden information, no adaptive adversary that can breach the bounds of the rules of the game. Given that, either brute force or ML/statistical approaches works well enough to build things like Deep Blue or the Google self-driving car.

At Palantir (where I work), we have, to date, stayed away from heavy machine learning or algorithmic approaches to data analysis, focusing instead on building better and better tools to connect human minds to data in a way that's rigorous and interactive. This is the only current way to detect adaptive adversaries like sophisticated fraudsters or state-sponsored cyber attackers - traditional ML approaches fail as tactics adapt faster than training data can be identified, tagged, and learned. I like to think of this class of problems as

arms races, since automation is easily defeated by a change in tactics that use more advanced/harder to detect techniques.

Let's take a look at fraud. Simple fraud, like stolen credit cards, is solved with simple automation looking for anomalies inside a big, validated training set (your legitimate transactions). The type of fraud is much more subtle I'm talking about goes something like this:

1. Open accounts under fake identities
2. Use accounts, pay balances, drive up credit lines
3. Hit magic number, max out all cards
4. Write fake checks to zero balances
5. Max cards a second time before the checks bounce
6. abandon the fake identities
7. Goto step 1

Instead, the state of the art is to use fairly simple data-mining to flag suspicious events with a low-threshold - yielding a set of candidate events much smaller than the initial haystack but containing a relatively large number of false positives. In the above example, we use clustering that looks at caller ID data for calls to the bank, card transactions, payment methods, IP address data for access to the website, and account details. Seemingly unrelated accounts that are linked are scored by aggregate credit risk and queued up for human analysis in a rich, interactive analytic environment. This is how the final determination of fraud is made. The magic of our software is not magic at all - we integrate the data across multiple, disparate sources of data into a human-conceptual model based on a constrained ontology of the problem at hand. The interface is interactive (sometimes requiring some pretty serious engineering to pull off) and speaks a language that is familiar to the human experts on the problem, enabling them to move through large volumes of data very quickly. (For a simple but demonstrative workflow, take a look at this cybersecurity demo)

Our newest innovation is letting the human analysts specify the pattern matching algorithms in this ontologically-typed language (Conceptually: "two credit card accounts used at the same store to buy the same item within three minutes of each other") that is then translated into search jobs on the map-reduce cluster. This allows for a tight feedback loop and hopefully stay ahead of the fraudsters.

What I've been noodling on is using a mix of MDF (these institutions fighting fraud are often combing through terabytes of data produced per day) and deep learning to see if you could train classifiers that would not only spot fraud but be able to see it as their tactics change.

It's sort of an interesting twist on the Turing Test: can an AI detect the signature of a human that's actively trying to evade detection?

--