From: Joscha Bach [redacted]>
To: Ari Gesher [redacted]
Cc: [redacted]>, Joi Ito <[redacted]>, takashi ikegami [redacted], Kevin Slavin <[redacted]>, Martin Nowak [redacted]>, Greg Borenstein <[redacted]>, Jeffrey Epstein <[redacted]>
Subject: Re: MDF
Date: Sun, 27 Oct 2013 01:47:29 +0000

Am 24.10.2013 um 02:10 schrieb Ari Gesher [redacted]

> 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.

There are a lot of games that do involve deception (for instance, most strategy games). At its core, deception is the intentional installation of a misleading belief in the mind of an adversary, usually to shape the adversary's actions in a way that increases the estimated likelihood of the realization of the goals of the instigator. Deception is a necessary by-product of the intelligence of a goal-directed system that is able to form a "theory of mind“ (i.e., a representation of the contents of the minds of others). Humans and other social animals do not need a special "deception module“, but built-in safeguards against runaway deception, so cooperation becomes possible.

> 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 (…) 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.
>
> (…)
>
> 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.

You make a very compelling point for the study of deception in the context of MDF (wrt credit card fraud). This is a very special case, of course: the situation you describe banks on deceiving an algorithm, rather than conning a human bank teller; the latter would involve very different strategies, ones that might be hard to detect statistically. Paradoxically, perhaps the latter might be easier to handle automatically, because most of the routines in human conmanship are built on human psychology. Fraud in electronic payment systems should be full of novelty and creativity, I imagine.
At some point, the electronic detection might become exhaustive. Instead of black-listing suspicious patterns and marking them for a human supervisor, it could model the individual lives and the likely interactions of the legitimate credit card holders to the point where it might be able to whitelist all non-suspicious activity. Before we have Strong AI, the application will still remain a hybrid between the intelligence of a human supervisor and the software (unless one wants to go the Paypal route and punish legitimate users whenever they trigger a false positive in the heuristics). At some point, one might want to call up the person to check up on their most recent purchase.
I do not understand yet why you are using so little machine learning at this point. Would it not help to cluster users, for instance, and detect if someone leaves their cluster? That might be hard to describe in an ontology!

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?

I doubt it… if the human actively tries to behave non-suspiciously, he may have an easier time to avoid the tripwires than a normal customer. He may also include the avoidance of "too good to be true" behavior. But what about a test for creativity and novelty?

Cheers,

Joscha