Avoiding a Volkswagen-style scandal in Artificial Intelligence

Originally published June 28, 2017 at https://www.aiia.net/artificial-intelligence/articles/avoiding-a-volkswagen-style-scandal-in-artificial

The boom of interest in Artificial Intelligence may not survive a major benchmark scandal

In June of 2015 a computer vision research team from Chinese web services giant Baidu caused scandal in the world of Machine Learning when it was revealed they had 'creatively' interpreted the rules of the world's most important computer vision research competition to gain competitive advantage.

The loophole in the regulations of the ImageNet Large Scale Visual Recognition Competition was semantic, a result of imperfect phrasing and the way English (the lingua franca of the global science community) re-uses the plural second-person personal pronoun 'you'.

The official rules stated 'Please note that you cannot make more than 2 submissions per week.'

The Baidu team, later claiming to have understood 'you' as addressing individual team members (rather than the more logical inference that it covered the entire contributing team), used multiple logins from team members to make multiple submissions, effectively gaming the terms of the challenge.

Baidu's claim of misinterpreting the rules was widely dismissed as disingenuous. All indications suggest that as the frameworks for Machine Learning and Artificial Intelligence solidify and new standards form, the inevitable benchmark tests which evaluate their progress will become subject to further attempts at corruption and deception.

There's massive precedent for this in the chequered history of GPU benchmarks for video cards - a prime market indicator for commercial advantage in the multi-billion dollar gaming industry.

Graphic lies

Though GPUs now have a direct relationship with AI research, the history of 'gaming' the gaming systems predates this association. In 2004 NVidia came under refreshed criticism for taking advantage of quirks in the 3Dmark2003 suite of graphic testing benchmarks in order to make its latest offering appear to perform better than previous outings. (Rival GPU manufacturer ATI is not able to offer a spotless record in this respect either.)

More recently Google Chrome deprecated the Octane JavaScript benchmark because of the avalanche of cheating that the standardised tests encouraged.

The company concluded that contestants were optimizing their submissions so tightly towards certain parameters of the benchmark that the code was actually less efficient in real world use than it might have been.

But the internet search giant doesn't hold the high ground in this regard either. In 2016 Amazon Web Services contested Google's claims that its CloudSQL database service had overcome a recent defeat in a performance show-down with Amazon Aurora by claiming that Google BigQuery beat Amazon Redshift in a subsequent test.

In the past few weeks a new release from mobile manufacturer OnePlus achieved instant notoriety when it was revealed that its software contains code designed to recognise up to seven testing suites and 'adapt' performance accordingly to achieve better results than the unit is likely to obtain in real world use.

But the biggest scandal of the last fifteen years around benchmark gaming is, naturally, the revelation that Volkswagen cheated on standardised software-based emissions tests over a period of many years — a news bomb which ultimately uncovered extensive industry gaming in the car manufacturing sector.

Inventing the game

At least the incidents of benchmark cheating in the GPU and car emissions sector are ultimately detectable, since the parameters are quite circumscribed; in the case of car manufacturers attempting to deceive about the levels of emissions their vehicles put out, the vehicle is programmed to behave differently when it detects test conditions. In the case of GPU cheats, the numerous possible 'shader dodges' (which in one case involved presenting a static instead of a dynamically redrawn image during a benchmark test) and other shallow illusions are set against known performance goals and strictly quantified performance parameters.

In the still-nascent field of AI research, the ground is incredibly fertile not only for gaming the tests, but to define the game — and all the rules surrounding it.

Cheatbots

Without necessarily intending to, the AI bubble has set itself up for future gaming scandals by being willing to benefit from the hype around chatbots, which are not really any kind of real Artificial Intelligence.

I've criticised previously the status of Chatbots as a 'litmus test' for the state of the art in Machine Learning. I can quite understand why facile systems such as Siri, Google Now and Alexa are becoming the popular poster-boys for AI.

I even acknowledge that they may be doing the sector a service in providing at least one focal and definitive offering in a field which — contrary to the way business likes its new prospects — is extraordinarily abstract and difficult to conceptualise.

But chatbots are in themselves a perfect example of 'gaming' expectations around Artificial Intelligence, without having any real input from neural networks underpinning their operation. Underpowered, easily befuddled, and never intended by their creator Alan Turing to constitute anything more than a thought-exercise around machine intellect, chatbots represent ersatz intelligence that uses the kind of limited data that you could probably extract from a human with a few web-cookies. They are *designed* to deceive, rather than to achieve.

So chatbots are probably not a template which AI should be too proud of as researchers into neural networks seek to develop genuinely evolutionary leaps in analytical power.

Nonetheless they are an obvious target for the kind of coding sleight-of-hand that can cause them to appear more interactive and knowledgeable about you than they can realistically be.

A cautionary note from history

The current state of interest in Artificial Intelligence and Machine Learning is practically a clone of the one that occurred in the 1960s in the wake of the perceptron and other novel developments around neural networks.

Now, as then, business is getting very excited about an economic sea-change in the making; now, as then, the news headlines are well-served by doomy tales of machine-driven apocalypse, bolstered by Hollywood output; now, as then, the academic and commercial research community is living in a methodical but excited fog of cross-fertilized ideas, eager investors, innovations and case studies.

But that 1960s furore devolved into the AI winter of the 1970s because neither the hardware nor the data infrastructure were remotely ready keep pace with the potential of the research.

As far as the hardware goes, that's still partially true - neural networks remain unsuitable candidates for low-powered mobile devices, and ruminative engines of invention in most test cases.

But this time we're a lot closer to that 1960s dream - useful neural networks under the attention of a far more connected global research community; networks which can provide actionable information on volumes of structured and unstructured data undreamed of fifty years ago.

But the nervousness about the prospect of a second AI winter seems likely to combine with the corporate sector's legendarily short-term thinking to invite new incidences of Barnum-style bluff - and a public disappointment around AI that could threaten the impetus of a movement that has genuine long-term validity, but still needs time to consolidate, retool and advance.

At the moment it's too fragile an ecostructure to hope to survive a major benchmarking scandal.

IMAGES: Flickr (License)