Artificial Intelligence: Threat or Menace?

By Charlie Stross

https://www.antipope.org/charlie/blog-static/2019/12/artificial-intelligence-threat.html

(*This is the text of a keynote talk I just delivered at the IT Futures conference held by the University of Edinburgh Informatics centre today.* NB: Some typos exist; I'll fix them tonight.)

Good morning. I'm Charlie Stross, and I tell lies for money. That is, I write fiction—deliberate non-truths designed to inform, amuse, and examine the human condition. More specifically, I'm a science fiction writer, mostly focusing on the intersection between the human condition and our technological and scientific environment: less Star Wars, more about bank heists inside massively multiplayer computer games, or the happy fun prospects for 3D printer malware.

One of the besetting problems of near-future science fiction is that life comes at you really fast these days. Back when I agreed to give this talk, I had no idea we'd be facing a general election campaign — much less that the outcome would already be known, with consequences that pretty comprehensively upset any predictions I was making back in September.

So, because I'm chicken, I'm going to ignore current events and instead take this opportunity to remind you that I can't predict the future. No science fiction writer can. Predicting the future isn't what science fiction is about. As the late Edsger Djikstra observed, "computer science is no more about computers than astronomy is about telescopes." He might well have added, or science fiction is about predicting the future. What I try to do is examine the human implications of possible developments, and imagine what consequences they *might* have. (Hopefully entertainingly enough to convince the general public to buy my books.)

So: first, let me tell you some of my baseline assumptions so that you can point and mock when you re-read the transcript of this talk in a decade's time.

Ten years in the future, we will be living in a world recognizable as having emerged from the current one by a process of continuous change. About 85% of everyone alive in 2029 is already alive in 2019. (Similarly, most of the people who're alive now will still be alive a decade hence, barring disasters on a historic scale.)

Here in the UK the average home is 75 years old, so we can reasonably expect most of the urban landscape of 2029 to be familiar. I moved to Edinburgh in 1995: while the Informatics Forum is new, as a side-effect of the disastrous 2002 old town fire, many of the university premises are historic. Similarly, half the cars on the road today will still be on the roads in 2029, although I expect most of the diesel fleet will have been retired due to exhaust emissions, and there will be far more electric vehicles around.

You don't need a science fiction writer to tell you this stuff: 90% of the world of tomorrow plus ten years is obvious to anyone with a weekly subscription to New Scientist and more imagination than a doorknob.

What's less obvious is the 10% of the future that isn't here yet. Of that 10%, you used to be able to guess most of it — 9% of the total — by reading technology road maps in specialist industry publications. We know what airliners Boeing and Airbus are starting development work on, we can plot the long-term price curve for photovoltaic panels, read the road maps Intel and ARM provide for hardware vendors, and so on. It was fairly obvious in 2009 that Microsoft would still be pushing some version of Windows as a platform for their hugely lucrative business apps, and that Apple would have some version of NeXTStep — excuse me, macOS — as a key element of their vertically integrated hardware business. You could run the same guessing game for medicines by looking at clinical trials reports, and seeing which drugs were entering second-stage trials — an essential but hugely expensive prerequisite for a product license, which requires a manufacturer to be committed to getting the drug on the market by any means possible (unless there's a last-minute show-stopper), 5-10 years down the line.

Obsolescence is also largely predictable. The long-drawn-out death of the pocket camera was clearly visible on the horizon back in 2009, as cameras in smartphones were becoming ubiquitous: ditto the death of the pocket GPS system, the compass, the camcorder, the PDA, the mp3 player, the ebook reader, the pocket games console, and the pager. Smartphones are technological cannibals, swallowing up every available portable electronic device that can be crammed inside its form factor.

However, this stuff ignores what Donald Rumsfeld named "the unknown unknowns". About 1% of the world of ten years hence always seems to have sprung fully-formed from the who-ordered-THAT dimension: we always get landed with stuff *nobody* foresaw or could possibly have anticipated, unless they were spectacularly lucky guessers or had access to amazing hallucinogens. And this 1% fraction of unknown unknowns regularly derails near-future predictions.

In the 1950s and 1960s, futurologists were obsessed with resource depletion, the population bubble, and famine: Paul Ehrlich and the other heirs of Thomas Malthus predicted wide-scale starvation by the mid-1970s as the human population bloated past the unthinkable four billion mark. They were wrong, as it turned out, because of the unnoticed work of a quiet agronomist, Norman Borlaug, who was pioneering new high yield crop strains: what became known as the Green Revolution more than doubled global agricultural yields within the span of a couple of decades. Meanwhile, it turned out that the most effective throttle on population growth was female education and emancipation: the rate of growth has slowed drastically and even reversed in some countries, and WHO estimates of peak population have been falling continuously as long as I can remember. So the take-away I'd like you to keep is that the 1% of unknown unknowns are often the most significant influences on long-term change.

If I was going to take a stab at identifying a potential 1% factor, the unknown unknowns that dominate for the second and third decade of the 21st century, I wouldn't point to climate change — the dismal figures are already quite clear — but to the rise of algorithmically targeted

advertising campaigns combined with the ascendancy of social networking. Our news media, driven by the drive to maximize advertising click-throughs for revenue, have been locked in a race to the bottom for years now. In the past half-decade this has been weaponized, in conjunction with data mining of the piles of personal information social networks try to get us to disclose (in the pursuit of advertising bucks), to deliver toxic propaganda straight into the eyeballs of the most vulnerable — with consequences that are threaten to undermine the legitimacy of democratic governmance on a global scale.

Today's internet ads are qualitatively different from the direct mail campaigns of yore. In the age of paper, direct mail came with a steep price of entry, which effectively limited it in scope — also, the print distribution chain was it relatively easy to police. The efflorescence of spam from 1992 onwards should have warned us that junk information drives out good, but the spam kings of the 1990s were just the harbinger of today's information apocalypse. The cost of pumping out misinformation is frighteningly close to zero, and bad information drives out good: if the propaganda is outrageous and exciting it goes viral and spreads itself for free.

The recommendation algorithms used by YouTube, Facebook, and Twitter exploit this effect to maximize audience participation in pursuit of maximize advertising click-throughs. They promote popular related content, thereby prioritizing controversial and superficially plausible narratives. Viewer engagement is used to iteratively fine-tune the selection of content so that it is more appealing, but it tends to trap us in filter bubbles of material that reinforces our own existing beliefs. And bad actors have learned to game these systems to promote dubious content. It's not just Cambridge Analytica I'm talking about here, or allegations of Russian state meddling in the 2016 US presidential election. Consider the spread of anti-vaccination talking points and wild conspiracy theories, which are no longer fringe phenomena but mass movements with enough media traction to generate public health emergencies in Samoa and drive-by shootings in Washington DC. Or the spread of algorithmically generated knock-offs of children's TV shows proliferating on YouTube that caught the public eye last year.

... And then there's the cute cat photo thing. If I could take a time machine back to 1989 and tell an audience like yourselves that in 30 years time we'd all have pocket supercomputers that place all of human knowledge at our fingertips, but we'd mostly use them for looking at kitten videos and nattering about why vaccination is bad for your health, you'd have me sectioned under the Mental Health Act. And you'd be acting reasonably by the standards of the day: because unlike fiction, emergent human culture is under no obligation to make sense.

Let's get back to the 90/9/1 percent distribution, that applies to the components of the near future: 90% here today, 9% not here yet but on the drawing boards, and 1% unpredictable. I came up with that rule of thumb around 2005, but the ratio seems to be shifting these days. Changes happen faster, and there are more disruptive unknown-unknowns hitting us from all quarters with every passing decade. This is a long-established trend: throughout most of recorded history, the average person lived their life pretty much the same way as their parents and grandparents. Long-term economic growth averaged less than 0.1% per year over the past two thousand years. It has only been since the onset of the industrial revolution that change has become a dominant influence on human society. I suspect the 90/9/1 distribution is now something more like 85/10/5 — that is, 85% of the world of 2029 is here today, about 10% can be anticipated, and the

random, unwelcome surprises constitute up to 5% of the mix. Which is kind of alarming, when you pause to think about it.

In the natural world, we're experiencing extreme weather events caused by anthropogenic climate change at an increasing frequency. Back in 1989, or 2009, climate change was a predictable thing that mostly lay in the future: today in 2019, or tomorrow in 2029, random-seeming extreme events (the short-term consequences of long-term climactic change) are becoming commonplace. Once-a-millennium weather outrages are already happening once a decade: by 2029 it's going to be much, *much* worse, and we can expect the onset of destabilization of global agriculture, resulting in seemingly random food shortages as one region or another succumbs to drought, famine, or wildfire.

In the human cultural sphere, the internet is pushing fifty years old, and not only have we become used to it as a communications medium, we've learned how to second-guess and game it. 2.5 billion people are on Facebook, and the internet reaches almost half the global population. I'm a man of certain political convictions, and I'm trying very hard to remain impartial here, but we have just come through a *spectacularly* dirty election campaign in which home-grown disinformation (never mind propaganda by external state-level actors) has made it almost impossible to get trustworthy information about topics relating to party policies. One party renamed its Twitter-verified feed from its own name to **FactCheckUK** for the duration of a televised debate. Again, we've seen search engine optimization techniques deployed successfully by a party leader — let's call him Alexander de Pfeffel something-or-other — who talked at length during a TV interview about his pastime of making cardboard model coaches. This led Google and other search engines to downrank a certain referendum bus with a promise about saving £350M a week for the NHS painted on its side, a promise which by this time had become deeply embarrassing.

This sort of tactic is viable in the short term, but in the long term is incredibly corrosive to public trust in the media — in all media.

Nor are the upheavals confined to the internet.

Over the past two decades we've seen revolutions in stock market and forex trading. At first it was just competition for rackspace as close as possible to the stock exchange switches, to minimize packet latency — we're seeing the same thing playing out on a smaller scale among committed gamers, picking and choosing ISPs for the lowest latency — then the high frequency trading arms race, in which case fuzzing the market by injecting "noise" in the shape of tiny but frequent trades allowed volume traders to pick up an edge (and effectively made small-scale day traders obsolete). I lack inside information but I'm pretty sure if you did a deep dive into what's going on behind the trading desks at FTSE and NASDAQ today you'd find a *lot* of powerful GPU clusters running Generative Adversarial Networks to manage trades in billions of pounds' worth of assets. Lights out, nobody home, just the products of the post-2012 boom in deep learning hard at work, earning money on behalf of the old, slow, procedural AIs we call corporations.

What do I mean by that — calling corporations AIs?

Although speculation about mechanical minds goes back a lot further, the field of Artificial Intelligence was largely popularized and publicized by the groundbreaking 1956 Dartmouth Conference organized by Marvin Minsky, John McCarthy, Claude Shannon, and Nathan Rochester of IBM. The proposal for the conference asserted that, "every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it", a proposition that I think many of us here would agree with, or at least be willing to debate. (Alan Turing sends his apologies.) Furthermore, I believe mechanisms exhibiting many of the features of human intelligence had already existed for some centuries by 1956, in the shape of corporations and other bureaucracies. A bureaucracy is a framework for automating decision processes that a human being might otherwise carry out, using human bodies (and brains) as components: a corporation adds a goal-seeking constraints and real-world i/o to the procedural rules-based element.

As justification for this outrageous assertion — that corporations are AIs — I'd like to steal philosopher John Searle's "Chinese Room" thought experiment and misapply it creatively. Searle, a skeptic about the post-Dartmouth Hard AI project — the proposition that symbolic computation could be used to build a mind — suggested the thought experiment as a way to discredit the idea that a digital computer executing a program can be said to have a mind. But I think he inadvertently demonstrated something quite different.

To crib shamelessly from wikipedia:

Searle's thought experiment begins with this hypothetical premise: suppose that artificial intelligence research has constructed a computer that behaves as if it understands Chinese. It takes Chinese characters as input and, by following the instructions of a computer program, produces other Chinese characters, which it presents as output. Suppose, says Searle, that this computer comfortably passes the Turing test, by convincing a human Chinese speaker that the program is itself a live Chinese speaker. To all of the questions that the person asks, it makes appropriate responses, such that any Chinese speaker would be convinced that they are talking to another Chinese-speaking human being.

The question Searle asks is: does the machine literally "understand" Chinese? Or is it merely simulating the ability to understand Chinese?

Searle then supposes that he is in a closed room and has a book with an English version of the computer program, along with sufficient papers, pencils, erasers, and filing cabinets. Searle could receive Chinese characters through a slot in the door, process them according to the program's instructions, and produce Chinese characters as output. If the computer had passed the Turing test this way, it follows that he would do so as well, simply by running the program manually.

Searle asserts that there is no essential difference between the roles of the computer and himself in the experiment. Each simply follows a program, step-by-step, producing a behavior which is then interpreted by the user as demonstrating intelligent conversation. But Searle himself would not be able to understand the conversation.

The problem with this argument is that it is apparent that a company is nothing but a very big Chinese Room, containing a large number of John Searles, all working away at their rule sets and inputs. We many not agree that an AI "understands" Chinese, but we can agree that it performs symbolic manipulation; and a room full of bureaucrats looks awfully similar to a hypothetical Turing-test-passing procedural AI from here.

Companies don't literally try to pass the Turing test, but they exchange information with other companies — and they are powerful enough to process inputs far beyond the capacity of an individual human brain. A Boeing 787 airliner contains on the order of six million parts and is produced by a consortium of suppliers (coordinated by Boeing); designing it is several orders of magnitude beyond the competence of any individual engineer, but the Boeing "Chinese Room" nevertheless developed a process for designing, testing, manufacturing, and maintaining such a machine, and it's a process that is not reliant on any sole human being.

Where, then, is Boeing's mind?

I don't think Boeing has a mind *as such*, but it functions as an ad-hoc rules-based AI system, and exhibits drives that mirror those of an actual life form. Corporations grow, predate on one another, seek out sources of nutrition (revenue streams), and invade new environmental niches. Corporations exhibit metabolism, in the broadest sense of the word — they take in inputs and modify them, then produce outputs, including a surplus of money that pays for more inputs. Like all life forms they exist to copy information into the future. They treat human beings as interchangeable components, like cells in a body: they function as superorganisms — hive entities — and they reap efficiency benefits when they replace fallible and fragile human components with automated replacements.

Until relatively recently the automation of corporate functions was limited to mid-level bookkeeping operations — replacing ledgers with spreadsheets and databases — but we're now seeing the spread of robotic systems outside manufacturing to areas such as lights-out warehousing, and the first deployments of deep learning systems for decision support.

I spoke about this at length a couple of years ago in a talk I delivered at the Chaos Communications Congress in Leipzig, titled "Dude, You Broke the Future" — you can find it on YouTube and a text transcript on my blog — so I'm not going to dive back into that topic today. Instead I'm going to talk about some implications of the post-2012 AI boom that weren't obvious to me two years ago.

Corporations aren't the only pre-electronic artificial intelligences we've developed. Any bureaucracy is a rules-based information processing system. Governments are superorganisms that behave like very large corporations, but differ insofar as they can raise taxes (thereby creating demand for circulating money, which they issue), stimulating economic activity. They can recirculate their revenue through constructive channels such as infrastructure maintenance, or destructive ones such as military adventurism. Like corporations, governments are potentially immortal until an external threat or internal decay damages them beyond repair. By promulgating and enforcing laws, governments provide an external environment within which the much smaller rules-based corporations can exist.

(I should note that at this level, it doesn't matter whether the government's claim to legitimacy is based on the will of the people, the divine right of kings, or the Flying Spaghetti Monster: I'm talking about the mechanical working of a civil service bureaucracy, *what* it does rather than *why* it does it.)

And of course this brings me to a third species of organism: academic institutions like the University of Edinburgh.

Viewed as a corporation, the University of Edinburgh is impressively large. With roughly 4000 academic staff, 5000 administrative staff, and 36,000 undergraduate and postgraduate students (who may be considered as a weird chimera of customers and freelance contractors), it has a budget of close to a billion pounds a year. Like other human superorganisms, Edinburgh University exists to copy itself into the future — the climactic product of a university education is, of course, a professor (or alternatively a senior administrator), and if you assemble a critical mass of lecturers and administrators in one place and give them a budget and incentives to seek out research funding and students, you end up with an academic institution.

Quantity, as the military say, has a quality all of its own. Just as the Boeing Corporation can undertake engineering tasks that dwarf anything a solitary human can expect to achieve within their lifetime, so too can an institution out-strip the educational or research capabilities of a lone academic. That's *why* we have universities: they exist to provide a basis for collaboration, quality control, and information exchange. In an idealized model university, peers review one another's research results and allocate resources to future investigations, meanwhile training undergraduate students and guiding postgraduates, some of whom will become the next generation of researchers and teachers. (In reality, like a swan gliding serenely across the surface of a pond, there's a lot of thrashing around going on beneath the surface.)

The corpus of knowledge that a student needs to assimilate to reach the coal face of their chosen field exceeds the competence of any single educator, so we have division of labour and specialization among the teachers: and the same goes for the practice of research (and, dare I say it, writing proposals and grant applications).

Is the University of Edinburgh itself an artificial intelligence, then?

I'm going to go out on a limb here and say " not *yet*". While the University Court is a body corporate established by statute, and the administration of any billion pound organization of necessity shares traits with the other rules-based bureaucracies, we can't reasonably ascribe a theory of mind, or actual self-aware consciousness, to a university. Indeed, we can't ascribe consciousness to *any* of the organizations and processes around us that we call AI.

Artificial Intelligence really has come to mean three different things these days, although they all fall under the heading of "decision making systems opaque to human introspection". We have the classical bureaucracy, with its division of labour and procedures executed by flawed, fallible human components. Next, we have the rules-based automation of the 1950s through 1990s, from Expert Systems to Business Process Automation systems — tools which improve the efficiency and reliability of the previous bureaucratic model and enable it to operate with fewer human cogs

in the gearbox. And since roughly 2012 we've had a huge boom in neural computing, which I guess is what brings us here today.

Neural networks aren't new: they started out as an attempt in the early 1950s to model the early understanding of how animal neurons work. The high level view of nerves back then — before we learned a lot of confusing stuff about pre- and post-synaptic receptor sites, receptor subtypes, microtubules, and so on — is that they're wiring and switches, with some basic additive and subtractive logic superimposed. (I'm going to try not to get sidetracked into biology here.) Early attempts at building recognizers using neural network circuitry, such as 1957's Perceptron network, showed initial promise. But they were sidelined after 1969 when Minsky and Papert formally proved that a perceptron was computationally weak — it couldn't be used to compute an Exclusive-OR function. As a result of this resounding vote of no-confidence, research into neural networks stagnated until the 1980s and the development of backpropagation. And even with a more promising basis for work, the field developed slowly thereafter, hampered by the then-available computers.

A few years ago I compared the specifications for my phone — an iPhone 5, at that time — with a Cray X-MP supercomputer. By virtually every metric, the iPhone kicked sand in the face of its 30-year supercomputing predecessor, and today I could make the same comparison with my wireless headphones or my wrist watch. We tend to forget how inexorable the progress of Moore's Law has been over the past five decades. It has brought us roughly ten orders of magnitude of performance improvements in storage media and working memory, a mere nine or so orders of magnitude in processing speed, and a dismal seven orders of magnitude in networking speed.

In search of a concrete example, I looked up the performance figures for the GPU card in the newly-announced Mac Pro; it's a monster capable of up to 28.3 Teraflops, with 1Tb/sec memory bandwidth and up to 64Gb of memory. This is roughly equivalent to the NEC Earth Simulator of 2002, a supercomputer cluster which filled 320 cabinets, consumed 6.4 MW of power, and cost the Japanese government 60 billion Yen (or about £250M) to build. The Radeon Pro Vega II Duo GPU I'm talking about is obviously much more specialized and doesn't come with the 700Tb disks or 1.6 petabytes of tape backup, but for raw numerical throughput — which is a key requirement in training a neural network — it's competitive. Which is to say: a 2020 workstation is roughly as powerful as half a billion pounds-worth of 2002 supercomputer when it comes to training deep learning applications.

In fact, the iPad I'm reading this talk from — a 2018 iPad Pro — has a processor chipset that includes a dedicated 8-core neural engine capable of processing 5 trillion 8-bit operations per second. So, roughly comparable to a mid-90s supercomputer.

Life (and Moore's Law) comes at you fast, doesn't it?

But the news on the software front is less positive. Today, our largest neural networks aspire to the number of neurons found in a mouse brain, but they're structurally far simpler. The largest we've actually trained to do something useful are closer in complexity to insects. And you don't

have to look far to discover the dismal truth: we may be able to train adversarial networks to recognize human faces most of the time, but there are also famous failures.

For example, there's the Home Office passport facial recognition system deployed at airports. It was recently reported that it has difficulty recognizing faces with very pale or very dark skin tones, and sometimes mistakes larger than average lips for an open mouth. If the training data set is rubbish, the output is rubbish, and evidently the Home Office used a training set that was not sufficiently diverse. The old IT proverb applies, "garbage in, garbage out" — now with added opacity.

The key weakness of neural network applications is that they're only as good as the data set they're trained against. The training data is invariably curated by humans. And so, the deep learning application tends to replicate the human trainers' prejudices and misconceptions.

Let me give you some more cautionary tales. Amazon is a huge corporation, with roughly 750,000 employees. That's a huge human resources workload, so they sank time and resources into training a network to evaluate resumes from job applicants, in order to pre-screen them and spit out the top 5% for actual human interaction. Unfortunately the training data set consisted of resumes from existing engineering employees, and even more unfortunately a very common underlying quality of an Amazon engineering employee is that they tend to be white and male. Upshot: the neural network homed in on this and the project was ultimately cancelled because it suffered from baked-in prejudice.

Google Translate provides is another example. Turkish has a gender-neutral pronoun for the third-person singular that has no English-language equivalent. (The closest would be the third-person plural pronoun, "they".) Google Translate was trained on a large corpus of documents, but came down with a bad case of gender bias in 2017, when it was found to be turning the neutral pronoun into a "he" when in the same sentence as "doctor" or "hard working," and a "she" when it was in proximity to "lazy" and "nurse."

Possibly my favourite (although I drew a blank in looking for the source, so you should treat this as possibly apocryphal) was a DARPA-funded project to distinguish NATO main battle tanks from foreign tanks. It got excellent results using training data, but wasn't so good in the field ... because it turned out that the recognizer had gotten very good at telling the difference between snow and forest scenes and arms trade shows. (Russian tanks are frequently photographed in winter conditions — who could possibly have imagined *that*?)

Which brings me back to Edinburgh University.

I can speculate wildly about the short-term potential for deep learning in the research and administration areas. Research: it's a no-brainer to train a GAN to do the boring legwork of looking for needles in the haystacks of experimental data, whether it be generated by genome sequencers or radio telescopes. Technical support: just this last weekend I was talking to a bloke whose startup is aiming to use deep learning techniques to monitor server logs and draw sysadmin attention to anomalous patterns in them. Administration: if we can just get past the "white, male" training trap that tripped up Amazon, they could have a future in screening job

candidates or student applications. Ditto, automating helpdesk tasks — the 80/20 rule applies, and chatbots backed by deep learing could be a very productive tool in sorting out common problems before they require human intervention. This stuff is obvious.

But it's glaringly clear that we need to get better — much better — at critiquing the criteria by which training data is compiled, and at working out how to sanity-test deep learning applications.

For example, consider a GAN trained to evaluate research grant proposals. It's almost inevitable that some smart-alec will think of this (and then attempt to use feedback from GANs to *improve* grant proposals, by converging on the set of characteristics that have proven most effective in extracting money from funding organizations in the past). But I'm almost certain that any such system would tend to recommend against ground-breaking research by default: promoting proposals that resemble past work research is no way to break new ground.

Medical clinical trials focus disproportionately on male subjects, to such an extent that some medicines receive product licenses without being tested on women of childbearing age at all. If we use existing trials as training data for identifying possible future treatments we'll inevitably end up replicating historic biases, missing significant opportunities to improve breakthrough healthcare to demographics who have been overlooked.

Or imagine the uses of GANs for screening examinations — either to home in on patterns indicative of understanding in essay questions (grading essays being a huge and tedious chore), or (more controversially) to identify cheating and plagiarism. The opacity of GANs means that it's possible that they will encode some unsuspected prejudices on the part of the examiners whose work they are being trained to reproduce. More troublingly, GANs are vulnerable to adversarial attacks: if the training set for a neural network is available, it's possible to identify inputs which will exploit features of the network to produce incorrect outputs. If a neural network is used to gatekeep some resource of interest to human beings, human beings will try to pick the lock, and the next generation of plagiarists will invest in software to produce false negatives when their essay mill purchases are screened.

And let's not even think about the possible applications of neurocomputing to ethics committees, not to mention other high-level tasks that soak up valuable faculty time. Sooner or later someone will try to use GANs to pre-screen proposed applications of GANs for problems of bias. Which might sound like a worthy project, but if the bias is already encoded in the ethics monitoring neural network, experiments will be allowed to go forward that really shouldn't, and vice versa.

Professor Noel Sharkey of Sheffield University went public yesterday with a plea for decision algorithms that impact peoples' lives — from making decisions on bail applications in the court system, to prefiltering job applications — to be subjected to large-scale trials before roll-out, to the same extent as pharmaceuticals (which have a similar potential to blight lives if they aren't carefully tested). He suggests that the goal should be to demonstrate that there is no statistically significant in-built bias before algorithms are deployed in roles that detrimentally affect human subjects: he's particularly concerned by military proposals to field killer drones without a human being in the decision control loop. I can't say that he's wrong, because he's very, very right.

"Computer says no" was a funny catch-phrase in "Little Britain" because it was really an excuse a human jobsworth used to deny a customer's request. It's a whole lot less funny when it really *is* the computer saying "no", and there's no human being in the loop. But what if the computer is saying "no" because its training data doesn't like left-handedness or Tuesday mornings? Would you even know? And where do you go if there's no right of appeal to a human being?

So where is AI going?

Now, I've just been flailing around wildly in the dark for half an hour. I'm probably laughably wrong about some of this stuff, especially in the detail level. But I'm willing to stick my neck out and make some firm predictions.

Firstly, for a decade now IT departments have been grappling with the bring-your-own-device age. We're now moving into the bring-your-own-neural-processor age, and while I don't know what the precise implications are, I can see it coming. As I mentioned, there's a neural processor in my iPad. In ten years time, future-iPad will probably have a neural processor three orders of magnitude more powerful (at least) than my current one, getting up into the trillion ops per second range. And all your students and staff will be carrying this sort of machine around on their person, all day. In their phones, in their wrist watches, in their augmented reality glasses.

The Chinese government's roll-out of social scoring on a national level may seem like a dystopian nightmare, but something not dissimilar could be proposed by a future university administration as a tool for evaluating students by continuous assessment, the better to provide feedback to them. As part of such a program we could reasonably expect to see ubiquitous deployment of recognizers, quite possibly as a standard component of educational courseware. Consider a distance learning application which uses gaze tracking, by way of a front-facing camera, to determine what precisely the students are watching. It could be used to provide provide feedback to the lecturer, or to direct the attention of viewers to something they've missed, or to pay for the courseware by keeping eyeballs on adverts. Any of these purposes are possible, if not desirable.

With a decade's time for maturation I'd expect to see the beginnings of a culture of adversarial malware designed to fool the watchers. It might be superficially harmless at first, like tools for fooling the gaze tracker in the aforementioned app into thinking a hung-over student is not in fact asleep in front of their classroom screen. But there are darker possibilities, and they only start with cheating continuous assessments or faking research data. If a future Home Office tries to automate the PREVENT program for detecting and combating radicalization, or if they try to extend it — for example, to identify students holding opinions unsympathetic to the governing party of the day — we could foresee pushback from staff and students, and some of the pushback could be algorithmic.

This is proximate-future stuff, mind you. In the long term, all bets are off. I am not a believer in the AI singularity — the rapture of the nerds — that is, in the possibility of building a brain-in-a-box that will self-improve its own capabilities until it outstrips our ability to keep up. What CS professor and fellow SF author Vernor Vinge described as "the last invention humans will ever need to make". But I *do* think we're going to keep building more and more complicated, systems

that are opaque rather than transparent, and that launder our unspoken prejudices and encode them in our social environment. As our widely-deployed neural processors get more powerful, the decisions they take will become harder and harder to question or oppose. And that's the real threat of AI — not killer robots, but "computer says no" without recourse to appeal.

I'm running on fumes at this point, but if I have any message to leave you with, it's this: AI and neurocomputing isn't magical and it's not the solution to all our problems, but it *is* dangerously non-transparent. When you're designing systems that rely on AI, please bear in mind that neural networks can fixate on the damndest stuff rather than what you want them to measure. Leave room for a human appeals process, and consider the possibility that your training data may be subtly biased or corrupt, or that it might be susceptible to adversarial attack, or that it turns yesterday's prejudices into an immutable obstacle that takes no account of today's changing conditions.

And please remember that the point of a university is to copy information into the future through the process of educating human brains. And the human brain is still the most complex neural network we've created to date.