Mixing human and computational belief

[cross-post from Medium]

This is part 1 of a talk I gave recently on hacking belief systems using old AI techniques.

I have a thing about belief systems. Not religious belief, but what it means to believe something, when accuracy does and doesn't matter, and why "strongly held belief" is a better concept to aim at than "true". It's something I've been thinking about for a while, and I believe (!) it's a conversation we need to have both as technologists who base many of our work decisions on beliefs, and often work on a belief-based system (the internet), and in the current climate of "fake news", uncertainty and deliberate disinformation.

I care about belief because I care a lot about autonomy: the ways that humans and machines can work together, sharing control, responsibility and as teams, and we don't talk enough yet about the human-machine sharing of belief systems and cultures that go with that. I also work in two professions that are fundamentally based on belief: data science, and lean/agile system development, and care about two topics that badly need a better theory of belief: that deep learning systems are doubling-down on human inequality and bias, and that America's collective belief systems are currently being 'hacked'. If that isn't enough, there's also a smaller interest, that Aspergers and neurotypical people hold the world in their minds differently, and whether that could be useful.

There are many definitions of belief, and many interpretations of those definitions. For this talk, I looked primarily at three old definitions: doxa, faith-based division of 'facts' into 'true' or 'false' (doxa is the root word of 'orthodoxy'); pistis, which is more nuanced, and allows for evidence and different degrees of confidence (pistis is the root word of epistemology), and normative belief: what you think other people expect you to believe. Each of these exist in both humans and systems, and each can be manipulated (albeit often in different ways).

You don't escape the machine

Robot image from clipart library http://clipart-library.com/clipart/pT7rMdb8c.htm

We need to find a way to frame human and machine belief in the same way, so we can talk about their joint effects and manipulations, and how to apply theories about one to the other. We can frame our discussion of human beliefs in the same way that we talk about robots; humans as self-contained units that rely entirely on their sensors (things that gather information about the world, e.g. eyes) and effectors (things that can change the state of the world, e.g. hands), and build belief-based models of their world based on those interactions.

There are many other philosophies about what it means to be human, but the one I'm holding here is that we don't escape the machine: we will probably never totally know our objective truths, we're probably not in the matrix, but we humans are all systems whose beliefs in the world are completely shaped by our physical senses, and those senses are imperfect. We'll rarely have complete information either (e.g. there are always outside influences that we can't see), so what we really have are very strong to much weaker beliefs.

There are some beliefs that we accept as truths (e.g. I have a bruise on my leg because I walked into a table today), but mostly we're basing what we believe on a combination of evidence and personal viewpoint, e.g. "it's not okay to let people die because they don't have healthcare", where that personal viewpoint is formed from earlier evidence and learning.

First technique: uncertain reasoning

So now we have humans and bots framed the same way, let's look at some of the older 'bot reasoning systems that might help. First up are belief networks (also known as Bayesian networks, probabilistic networks etc).

Image: wikipedia article on belief networks

Let’s look at the structure of these networks first. This is the classic tiniest network: a way of talking about the interactions between wet grass, rainfall and a sprinkler system (which may or may not be on). The arrows (and not every probabilistic network had arrows) show causality: if it rains, the grass gets wet; if the sprinkler is on, the grass gets wet, if it rains, the sprinkler is less likely to be on. And with this tiny network, we can start to talk about interactions between beliefs that don’t just follow the arrows, e.g. if the grass is wet, why?

This is how we used to do things before the explosion of data that came with the Internet, and everyone and everything generating data across it. Artificial Intelligence was about reasoning systems, where we tried to replicate either experts (e.g. “expert systems”) or the beliefs than an average rational person would hold (e.g. “normative belief”). Without enough data to throw at a deep learning system and allow the system to work out connections, we focussed on building those connections ourselves. And frankly, most expert systems were pretty dumb (“if x then y” repeated, usually with a growing morass of edge cases); the interesting work happened where people were thinking about how to handle uncertainty in machine reasoning.

Image: wikipedia article on belief networks

Here’s that network again, but in its full form, with the interactions between nodes. Each of those links applies Bayes’ theory to describe how our belief in the state of the thing at the end of a set of links, e.g. wet grass, depends on the states of the things at the starts of the links, e.g. sprinkler and rain (aside: I first learnt about this network on the week that I also learnt that running across an American lawn at 7am in summertime could get you soaked). These things chain together, so for instance if I have a weather forecast, I can give a belief about the (future) grass being wet via my belief in (future) rain.

Back in the day there were many uncertain reasoning techniques that we could slot into networks like this: Bayesian probability theory, multi-state logics (e.g. (true, false, unknown) as three possible values for a ‘fact’), modal logics (reasoning about ‘necessary’ and ‘possible’), interval probabilities, fuzzy logic, possibility theory, interval probabilities, but eventually Bayesian probability theory became dominant in the way we thought about probability (and as far as I know, is the only uncertainty theory taught on most data science courses beyond classical frequentist probability). These techniques have influenced other potentially useful tools like structured analytics, and it might be useful to reexamine some of them later.

Meanwhile, even a small network like this raises lots of questions: what about other reasons for wet grass, what if we’re wrong, what if it’s a really important decision (like surgery) etc etc etc. There are framings for these questions too (the frame problem, risk-averse reasoning etc), that could be useful for one of my starting problems: deep learning systems doubling-down on human inequality and bias.

Learning = forming beliefs

Images: two recent headlines about systems that may/do have human bias built into them.

We've talked a bit about human belief, and about machine representations of uncertainty and belief. The next part of the human-bot connection is to explore how computers have beliefs too, how we're not thinking about those belief sets in the same way that we think about human beliefs, and about how we really should start to do that.

We train our algorithms like we train our children: in most machine learning, we give them examples of right and wrong; in some machine learning, we correct them when their conclusions aren't sound. Except we don't do all the things that we do with children when we train a machine: with children, we install culture, correct 'errors' and deviations from that, and parents from the pistis traditions do many other things to help children question their environment and conclusions.

And this lack of the other things is starting to show, and really starting to matter, whether it be researchers touting their great classification scores without wondering if there's a racial bias in sentencing in China, chatbots mirroring our worst online selves, or race bias being built into government systems that make decisions about people's lives from education to health to prison sentences all over the USA. We're starting to talk about the problems here, but we also need to talk about the roots and mitigations of those problems, and some of those might come from asking what the algorithmic equivalent of a good childhood learning is. I don't have a good answer for that yet, but gut feel is that autonomy theory could be part of that response.

Beliefs can be shared

Image: clipart library http://clipart-library.com/clipart/pT7rMdb8c.htm

In machine learning we're often talking about a single set of beliefs from a single viewpoint as a good representation of the world. When we look at humans, we talk a lot more about other peoples' beliefs, and the interactions between their belief systems and viewpoints.

When we talk about belief sharing, we're talking about communication: the ways in which we pass or share beliefs. And communication is totally and routinely hackable, either in-message through the messages used, messages adapted, subtexts created, or by adapting communications channels and using old-school network effects like simple and complex contagion.

As technologists, we spend a lot of our time talking about other people's beliefs. But we also need to talk about what it means to understand what someone else's beliefs are. When we talk about other people's beliefs, we're really talking about our beliefs about their beliefs which may be offered through a communication channel (e.g. surveys) which is subject to their beliefs about our beliefs and how we're likely to react to them. This is important, complicated, and something that I had many fascinating meetings about when I worked on intelligence analysis… which also has a set of techniques that could help.

PACT image from book Human Factors in Simulation and Training

https://books.google.com/books?id=cgT56UW6aPUC

Humans and machines can share beliefs too. There's a whole field of study and practice (autonomy theory) on how machines (originally UAVs) and humans share control, authority, viewpoint and communication. PACT is one of the early models — a relative youngster at about 15–20 years old; with the rise of things like self-driving cars, new models mirroring this work are appearing, but although there are discussions about sensing and action, there hasn't been much yet on shared belief and culture.

Yes, culture, because when you mix humans and bots in UAV teams, the same cultural issues that happen when you mix, say, Italian and American pilots in a team, start to happen between the bots and the humans: the bots' social rules differ, their responses differ and building shared mental models can become strained. When we look at the internet as a mixed human-machine belief system, this earlier work might become important.

Beliefs can be hacked

Text from an article about Cambridge Analytica's work on Trump's election

campaign. https://www.theguardian.com/politics/2017/feb/26/robert-mercer-breitbart-war-on-media-steve-bannon-donald-trump-nigel-farage

You can hack computer beliefs; you can also hack human beliefs. Here's a quote from a recent article about Trump's election campaign. This looks suspiciously like a computer virus infecting and being carried by humans in their belief systems, with network contagion and adaptation (spreading and learning). We have a crossover with the bots too, with sensors (Facebook likes) and effectors (adverts): this takes the work beyond the simple network effects of memes and phemes (memes carrying false information).

Cambridge Analytica's system was large but apparently relatively simple, from the outside looking like an adapted marketing system; mixing that with belief and autonomy theories could produce something much more effective, albeit very ethically dubious.

Image: wikipedia page on the Muller-Lyer illusion

Even when you know better, it's hard to manage the biases created by someone hacking your belief system. This is a classic optical illusion (Muller-Lyer); I know that the horizontal lines are the same length, I've even measured them, but even knowing that, it's still hard to get my brain to believe it.

Optical illusions are a fun brain hack, and one that we can literally see, but there are subtle and non-visual brain hacks that are harder to detect and counter. We'll look at some of those later, but one insidious one is the use of association and memory traces: for example, if I borrow your car, tell you as a joke that I've crashed it, and immediately tell you it's a joke, the way that your brain holds new information (by first importing it as true and then denying it — see Daniel Gilbert's work for more on this) means that you not only keep a little trace of me having crashed your car in your memory, you also keep associated traces of me being an unsafe driver and are less likely to lend me your car again. This is a) why I don't make jokes like that, and b) the point of pretty much any piece of advertising content ever ("you're so beautiful if you have this…").

The Internet is made of belief

websearch result for "fake news"

The internet is made of many things: pages and and comment boxes and ports and protocols and tubes (for a given value of ‘tubes’). But it’s also made of belief: it’s a virtual space that’s only tangentially anchored in reality, and to navigate that virtual space, we all build mental models of who is out there, where they’re coming from, who or what to trust, and how to verify that they are who they say they are, and what they’re saying is true (or untrue but entertaining, or fantasy, or… you get the picture).

Location independence makes verification hard. The internet is (mostly) location-independent. That affects perception: if I say my favorite color is green, then the option to physically follow me and view the colours I like is only available to a few people; others must either follow my digital traces and my friends’ traces (which can be faked), believe me or decide to hold an open mind until more evidence appears.

When the carrier that we now use for many of our interactions with the world is so easily hacked (for both human and computer beliefs), we need to start thinking of counters to those hacks. These notes are a start at trying to think about what those counters might sensibly be.