What goes wrong in belief-based models?

[cross-post from Medium]

This is part 3 of a talk I gave recently on hacking belief systems using old AI techniques. Parts 1 and 2 looked at human and computational belief, and the internet and computer systems development as belief-based systems. Now we look at the things that can (and do) go wrong with belief-based systems.

Machine Belief

Machines have a belief system either because we hard-wire that system to give a known output for a known input (all those if-then statements), or because they learn it from data. We touched on machine learning in part 1 of this series (remember the racist chatbot?), and on how it's missing the things like culture, error-correction and questioning that we install in small humans. But before we get to how we might build those in, let's look at the basic mechanics of machine learning.

At the heart of machine learning are algorithms: series of steps that, repeated, create the ability to cluster similar things together, classify objects, and build inference chains. Algorithms are hackable too, and three big things tend to go wrong with them: inputs, models and willful abuse. Bias and missing data are issues on the input side: for instance, if an algorithm is fed racially-biased cause-to-effect inputs, it's likely to produce a similarly-biased system, and although we can interpolate between data points, if there's a demographic missing from our training data, we're likely to misrepresent them.

And although we may appear to have algorithms that automatically turn training data into models, there are almost always human design decisions being used to tune them, from the interpretations that we make of inputs (e.g. mouseclicks are equivalent to interest in a website, or that we make assumptions like 1 phone per person in a country means that everyone has a phone, when it could mean, as is common in West Africa, that some people have multiple phones and others have none) to the underlying mathematical models that they use.

Back in the day, we used to try to build models of the world by hand, with assertions like "all birds can fly", "birds have wings", "fred is a bird so fred can fly", and many arguments over what to do about penguins. That took time and created some very small and fragile systems, so now we automate a lot of that, by giving algorithms examples of input and output pairs, and leaving them to learn enough of a model that we get sensible results when we feed in an input it hasn't seen before. But sometimes we build systems whose models don't fit the underlying mechanics of the world that they're trying to capture, either because we're using the wrong type of model (e.g. trying to fit straight lines to a curve), the input data doesn't cover the things that we need it to learn, the world has changed since we built the model, or any of a dozen other things have gone wrong.

Machines can also be lied to. When we build models from data, we usually assume that the data, although inexact, isn't deliberately wrong. That's not true any more: systems can be 'gamed' with bad data (either misinformation, e.g. accidentally, or disinformation, e.g. deliberately), with machine learning from data on the internet being especially susceptible to this.

And issues in machine belief aren't abstract: as more systems are used to make decisions about people, Lazar's catalogue of algorithm risks become important to consider in building any algorithm (these risks include manipulation, bias, censorship, privacy violations, social discrimination, property rights violations, market power abuses, cognitive effects and heteronomy, e.g. individuals no longer having agency over the algorithms influencing their lives).

Human Belief

Humans and machines can go wrong in similar ways (missing inputs, incorrect assumptions, models that don't explain the input data etc), but the way that humans are wired adds some more interesting errors.

When we're shown a new fact (e.g. I crashed your car), our brains initially load in that fact as true, before we consciously decide whether it's true or not. That is itself a belief, but it's a belief that's consistent with human behaviour. And it has some big implications: if we believe, even momentarily, that something false is true, then by the time we reject it, it's a) left a trace of that belief in our brain, and b) our brain has built a set of linked beliefs (e.g. I'm an unsafe driver) that it doesn't remove with it. Algorithms can have similar memory trace problems; for instance, removing an input-output pair from a support vector machine is unlikely to remove its effect from the model, and unpicking bad inputs from deep learning systems is hard.

Humans have imperfect recall, especially when something fits their unconscious bias (try asking a US conservative about inauguration numbers). And humans also suffer from confirmation bias: they're more likely to believe something if it fits their existing belief framework. Humans also have mental immune systems, where new fact that are far enough away from their existing belief systems are immediately rejected. I saw this a lot when I worked as an innovations manager (we called this the corporate immune system): if we didn't introduce ideas very carefully and/or gradually, the cognitive dissonance in decision-makers' heads was too great, and we watched them shut those ideas out.

We're complicated. We also suffer from the familiarity backfire effect, as demonstrated by advertising: the more someone tries to persuade us that a myth is wrong, the more familiar we become with that myth, and more likely we are to believe it (familiarity and belief are linked). At the same time, we can also suffer from the overkill backfire effect, as demonstrated by every teenager on the planet ever, where the more evidence we are given against a myth, and the more complex that evidence is compared to a simple myth, the more resistant we are to believing it (don't despair: the answer is to avoid focussing on the myth whilst focussing on more simple alternative

explanations).

And we're social: even when presented with evidence that could make us change our minds, our beliefs are often normative (the beliefs that we think other people expect us to believe, often held so we can continue to be part of a group), and resistance to new beliefs is part of holding onto our in-group identity. Counters to this are likely to include looking at group dynamics and trust, and working out how to incrementally change beliefs not just at the individual but also at the group level.

Combined belief

Image: http://users.cs.cf.ac.uk/Dave.Marshall/AI2/node145.html

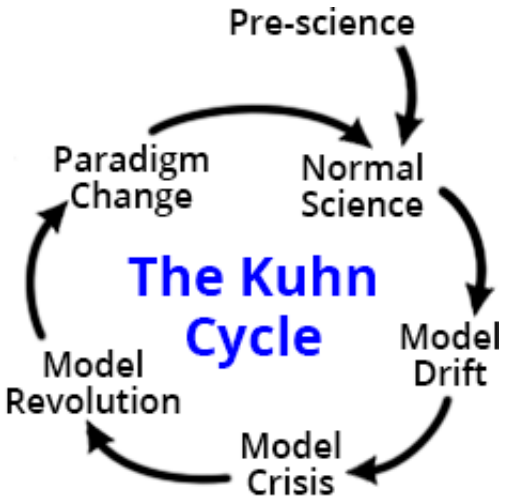
Human and machine beliefs have some common issues and characteristics. First, they have to deal with the frame problem: you can't measure the whole world, so there are always parts of your system potentially affected by things you haven't measured. This can be a problem; we spent a lot of the early 80s reasoning about children's building blocks (Winston's blocks world) and worrying about things like the "naughty baby problem", where sometime during our careful reasoning about the world, an outside force (the "naughty baby") could come into the frame and move all the block orders around.

We also have input issues, e.g. unknown data and uncertainty about the world. Some of that uncertainty is in the data itself; some of it is in the sources (e.g. other people) and channels (e.g. how facts are communicated); a useful trick from intelligence system design is to assess and score those 3 sources of uncertainty separately. Some uncertainty can be reduced by averaging: for instance, if you ask several farmers for the weight of a bull, each individual guess is likely to be wrong, but an average of the guesses is usually close to the real weight (this is how several cool tricks in statistics work).

And then there's human bias. A long time ago, if you wanted to study AI, you inevitably ended up reading up on cognitive psychology. Big data didn't exist (it did, but not applied to everything: it was mainly in things like real-time sonar processing systems), the Cyc system was a Sisyphean attempt to gather enough 'facts' to make a large expert system, and most of our efforts revolved around how to make sense of the world given a limited number of 'facts' about it (and also how to make sense of the world, given a limited number of uncertain data points and connections between them). And then the internet, and wikipedia, and internet 2.0 where we all became content providers, and suddenly we had so much data that we could just point machine learning algorithms at it, and start finding useful patterns.

We relied on two things: having a lot of data, and not needing to be specific: all we cared about was that, on average, the patterns worked. So we could sell more goods because the patterns that we used to target the types of goods to the types of people most likely to buy them, with the types of prompts that they were most likely to respond to. But we forgot the third thing: that almost all data collection has human bias in it — either in the collection, the models used or the assumptions made when using them. And that is starting to be a real problem for both humans and machines trying to make sense of the world from internet inputs.

And sometimes, beliefs are wrong



Science is cool — not just because blowing things up is acceptable, but also because good science knows that it's based on models, and actively encourages the possibility that those models are wrong. This leads to something called the Kuhn Cycle: most of the time we're building out a model of the world that we've collectively agreed is the best one available (e.g. Newtonian mechanics), but then at some point the evidence that we're collecting begins to break that collective agreement, and a new stronger model (e.g. relativity theory) takes over, with a strong change in collective belief. It usually takes an intelligent group with much humility (and not a little arguing) to do this, so it might not apply to all groups.

RedCross t-shirt after 2010 Chile earthquake

And sometimes people lie to you. The t-shirt above was printed to raise funds for the Red Cross after the Chile 2010 earthquake. It was part of a set of tweets sent pretending to be from a trapped mother and child, and was believed by the local fire brigade, who instead found an unaffected elderly couple at that address. Most crisismapping deployments have very strong verification teams to catch this sort of thing, and have been dealing with deliberate misinformation like this during crises since at least 2010. There’s also been a lot of work on veracity in data science (which added Veracity as a fourth “V” to “Volume, Velocity, Variety”).

Fox News visualisation: not exactly untrue, but not emphasising the truth either

And sometimes the person lying is you: if you’re doing data science, be careful not to deceive people by mistake, e.g. by not starting your axes at zero (see above), you can give readers a false impression of how significant differences between numbers are.