Should medical artificial intelligence be made to take the hippocratic oath (‘first, do no harm’)?

Hi I’m David Wong. I’m a lecturer in Health Informatics at the university of Leeds, and my background is in biomedical signal processing and machine learning. I’ve also built CE-marked software for hospital trusts – if you’re interested in any of that sort of thing, do contact me either by email (this address is valid for another couple of months) or via twitter.

This lunchtime, I want to convince you of two things. First, that artificial intelligence, as it exists in the real world is actually not very intelligent compared to you or me. Second, that this is ok, so long as we understand its limitations.

As I’m about to move posts to become a lecturer in AI for healthcare, the idea of telling people that artificial intelligence doesn’t really exist is a bit like … but I hope you’ll bear with me, and I hope that it doesn’t put me out of a job that I haven’t even started yet.

**Slide**

The topic of AI is super popular right now. For those who are starting up companies, there’s evidence to suggest that just including AI in your sales pitch attracts 15 to 50 percent more funding than other technology firms. This figure comes from a report from MMC, a London-based venture capital firm. They also said that those who said they only 40% of those who said they were AI companies actually used any sort of artificial intelligence.

**Slide**

Within healthcare, our secretary of state, Matt Hancock, is a strong proponent of using Artificial Intelligence. He said this last year at the launch of the AI Sector Deal:

*Read slide*

And he’s made many similar statements about how AI and digital technology in general will revolutionise healthcare. But what does AI even mean – do company investors know? Does Matt Hancock even know?

**Slide**

When the man on the street talks about Artificial Intelligence, what are the things that come to mind?

If you watch science fiction on TV, you might think about the homicidal crazy computer HAL in a space odyssey, or perhaps the holographic doctor from star trek or the android robot Data if you are a little older. For those in the pixar generation, you might think of Wall-E, the lonely robot looking for love.

Whilst they aren’t real, these movies portray synthetic intelligent creatures – they reason about the world they live in, they make decisions, and they understand the consequences of their decisions to some degree. They have learning that can be applied generally – put into a new and unexpected situation, they adapt to their surroundings much like we would expect a child to. This is what researchers call artificial general intelligence – it is very much still in the realm of science fiction.

**Slide**

When it comes to artificial intelligence in the real world, you might remember a few recent examples from the last few years.

A few years ago, IBM built a computer program to play the quiz show Jeopardy! On the whole, it did exceedingly well, and beat the best human competitors.

Many years before that, IBM built their second iteration of a chess computer that went on to beat the world’s best player at the time, Gary Kasparov.

In the last month or two, Deepmind created a program to play a real time strategy computer game. Again, it was as good, if not better, than the best human players.

These are three popular examples in which the computer was better than their human counterpart. However, the type of artificial intelligence displayed here is very different from the movies. These AI programs are extremely good at solving individual problems that are well defined – the rules are set out ahead of time, and cannot be broken.

But is this really intelligence?

**Slide**

Let’s consider this in a bit more detail. Now, I assume that most of you didn’t have a mis-spent childhood like me whiling away hours playing chess. So, let’s think about a much simpler game, noughts and crosses, instead.

To create a computer program to play this game, I can do a few things. I could get a computer to understand that there are 9 squares, and that it wins by getting three in a row, and it loses if the opponent gets three in a row. I could get it to look ahead and think about the implications of its moves. “if I play here, and then my opponent plays here, then I should play here…” that sort of thing. We can do that, because the game is so simple.

**Slide**

However, we can bypass that entirely by just using a set of simple instructions for each move. The instructions are here, and they will mean that a computer program will never lose a game of noughts and crosses.

* Move 1: Go in a corner.
* Move 2: If the other player did not go there then go in the opposite corner to move 1. Otherwise go in a free corner.
* Move 3: If there are two Xs and a space in a line (in any order) then go in that space. Otherwise if there are two Os and a space in a line then go in that space. Otherwise go in a free corner.
* Move 4: If there are two Xs and a space in a line (in any order) then go in that space. Otherwise if there are two Os and a space in a line then go in that space. Otherwise go in a free corner.
* Move 5: Go in the free space

Remarkably, I don’t need to know anything about how to win the game. I have done something that appears intelligent, without truly understanding what is going on. I haven’t had to think ahead about what the other person might do, but I can simply follow a set of rules.

Actually, the same applies to chess, and one can make a pretty decent chess computer without any idea of strategy or tactics, but simply by counting up pieces.

**Slide**

We see then, that it is easy to be fooled into thinking that there is some sort of intelligent behaviour going on, when there really isn’t. In fact, there is a famous example in the history of healthcare in which this happens – a chatbot called Eliza. Eliza was developed by researchers in 1964. It was an early chatbot that was supposed to be a psychotherapist. A user could enter a statement, and it would appear that Eliza knew what was being said, and that it could respond intelligently.

The picture here shows how it looked like originally, whereas there’s an updated version on the right here.

Actually, all it did was to identify key words in a sentence. If it saw something like ‘alike’ or ‘same’ – it could always ask something like ‘in what way?’. It didn’t need to know the answer, it just knew that this was always an acceptable response. Similarly, when it couldn’t work out a key word, it could just say something like ‘please go on’.

In itself, it is very very simple – the creator made it to show how BAD computers were at understanding people, but he was surprised to find that many users were convinced that ELIZA was intelligent and understanding.

**Slide**

Real world artificial intelligence methods are a bit more complicated than a noughts and crosses program, but are a lot closer to that, than the movie versions of AI. In the vast majority of cases we use a set of methods called supervised machine learning.

This is very useful when we want a computer program to help us to make a decision. For instance, I might have a photo, and I want the computer to decide whether it is a picture of a cat, a dog, or a fish.

To do this successfully, the computer needs lots of examples of cats, dogs and goldfish. It also needs to know, ahead of time, what each picture is.

From there, we try and simplify the picture into a set of key distinguishing features. For instance, we might use the average colour of the photo.

**Slide**

If we do this, and arrange all the photos by the average colour, we can see that most of the goldfish end up together. If we get a new picture, we simply see where it fits compared to all the other pictures, and group it with its closest neighbours.

To do better, we add more and more features. This is basically how AI works. Until ten or so years ago, a human person would have to work out the best characteristics to use. In the last ten years, we have collected enough data in some instances that we do not need to do this, and the computer can learn directly from the picture without having to be guided.

**Slide**

We can use the same approach in healthcare, and it works very well for some things. Here are a couple of examples.

We can use supervised classification AI to tell the difference between breast cancer mammograms with cancerous tumours vs those with no lumps, vs those with benign lumps.

**Slide**

We can also use it to diagnose heart attacks from ECG signals, or predict depression from facebook posts. There is a whole range of problems that we can solve using this very simple idea.

So, what we see if that we can solve very specific challenging problems – problems that are even hard for clinical experts to get right, by effectively comparing our new images to a database of old images.

**Slide**

So, very simplistically, our AI compares a new picture to ones that it already knows about. Is this really intelligence? I would suggest not. However, there is no denying that it is practically very useful indeed.

By understanding how AI models work conceptually, we can start to think where it might go wrong – particularly in medicine. Here are two possible shortcomings in which artificial intelligence could make terrible decisions.

**Slide**

First, because AI is effectively comparing new data to a database of old data, it means it doesn’t really know what to do when something really unique comes along.

For instance, if I give my AI pictures of hospitals, it will do fine for hospitals that look quite similar. What is if I give it a picture of an upside-down hospital? Well, it has never seen an upside down hospital before, and so it will get the decision wrong. In fact, to avoid this problem in practice, AI system developers put in rotated copies of their training data.

**Slide**

Here’s an example in practice, in which objects have been rotated in 3D. We can get similar problems in which the AI doesn’t understand which part of the image is important, so it can end up classifying using the background, rather than the object of interest.

We would never dream of these sorts of things happening to us, because we have some knowledge about objects, and know that they can be turned upside down. We know what sky and the ground are, and appreciate how they interact, in a way that current AI models do not.

**Slide**

The second issue is related to the first. We can be reasonably confident that we can get pretty far if our database of examples is good enough. However, if the database itself can’t be trusted, then the results can’t be trusted. The old computer scientists have expressed this for years as ‘Garbage In, Garbage Out’.

This isn’t just the problem of people entering the wrong numbers into hospital systems, though that happens too of course. More often, it is due to the fact that the data themselves are biased in some way.

One recent example of this from healthcare is:

**Slide**

This – Caruana and their colleagues trained a AI model to predict death based on hospital data. They found that asthmatics were less likely to die from pneumonia than people without asthma.

That doesn’t make any intuitive sense to a clinician. If I have a disease that affects my airways, and then I make it worse with pneumonia, surely that will make things worse?

Well, they discovered that all asthmatics with pneumonia went straight to ICU for the highest level of care, making them less likely to die. In this case, the human process was masked in the data.

Slide

Another example was published in the BMJ last year. They showed, by just observing the test data from hospital patients, first, that having a lab test done meant you were more likely to die. Secondly, that the actual test result was less relevant than the time that the test was ordered in many cases.

Of course, it’s not that lab tests cause more people to die, but it’s simply that having a lab test at all means that the dr has already deemed it necessary and is worried about you. If you are having a test done at 4am, when you ought to be sleeping, then the dr must be really really worried. So, you see, the data we have isn’t just a lab test number, but it contains all this additional hidden information about what the dr is thinking.

This happens time and time again in healthcare, in which the type of data recorded is already biased by the human processes in place.

**Unlike the first problem, which can be managed with careful thought – there is no easy way around this second problem.** The data isn’t necessarily the issue, but understanding what the data means is something that no AI can help us with.

In fact, this is one of the fundamental arguments around AI at the moment. It is very very good at finding patterns in data, but AI isn’t really reasoning. It can show that asthma patients don’t die as much, and that those woken up at night for test die more, but can’t tell us why.

**Slide**

Over lunch, I asked whether medical AI should be made to take the Hippocratic oath. I hope that from this talk you’ll understand that the answer is a resounding NO, simply because even the current advances in AI are not reasoning about the world, but doing looking for patterns. Rather, a better question is whether those who implement AI ought to be accountable for their unintended consequences – though that is a talk for another day.

There’s an old adage in computer science that computers are not intelligent, but the very opposite. They are very stupid, need to be told exactly what to do, but they never make a mistake. AI in healthcare may, at its core, be stupid, but as long as it is useful, and as long as we appreciate its limitations, then it’s ok with me.