The dangers of AI in health care: risk homeostasis and automation bias

Dr Cosima Gretton

Exhibition Road in the London Borough of Kensington & Chelsea is home to the some of the world's greatest museums. But its also part of a recent experiment in urban design, called Shared Space. Along the length of the street, the lines between road and pavement have been dissolved: cars and pedestrians share the same space, road markings, traffic lights and pedestrian crossings have disappeared. By increasing uncertainty and perceived risk the idea is that drivers will reduce their speed resulting in a safer environment for both pedestrians and vehicles.

Exhibition Road. Image Credit: La Citta Vita, via Creative Commons

The approach is based on the theory of risk homeostasis, first proposed by the Dutch psychologist Gerald Wilde in the 1980s. The theory draws on the observation that when an activity is made safer, the rate of accidents often remains the same. Mandatory seat belts reduce the likelihood of an injury in an accident, but don't reduce the death rate per capita. Drivers with anti-lock brakes drive closer to the car in front. When in 1967 Sweden switched over to driving on the right, there was a marked reduction in the rate of fatalities, which returned to the original rate a year and a half later. Human risk taking behavior seems to be tightly coupled to the perceived level of danger. Reduce how risky the activity feels, and people will be more daring.

> "The greater the number of prescriptions, the more people's sense of personal responsibility dwindles." (Hans Monderman)

Risk homeostasis has been controversial since its inception, but over the last few decades the idea of behavioral adaptation to perceived risk has become accepted by the scientific community.

Putting it into practice seems to work in some cases. A Shared Space approach was implemented just around the corner from Exhibition Road, in High Street Kensington. Analysis of public data on the two years prior to the change and the two years subsequent showed a 43% reduction in traffic related injuries.

Human-human risk homeostasis

Risk in clinical practice is often obfuscated by the complexities of the science. But evidence of risk homeostasis between clinicians has been found, for example, in a recent study of nurses in an Intensive Care Unit in the UK. Safety measures implemented during drug dispensing involve multiple cross-checks by different colleagues before a drug is given to a patient. Although nurses are trained to double-check, the safety measures reduce the perceived level of risk and the nurses in this study assumed a mistake was less likely to be made.

> "I think the staff are very trusting of one another, when checking drugs, instead of looking carefully at the prescription as they should do. Umm, because they take it for granted that you wouldn't make a mistake I suppose". (Sanghera et. al, 2007)

In his book the Digital Doctor, Bob Wachter tells the story of Pablo Garcia, a young patient given a 38 fold overdose of an anti-epileptic. He describes how staff failed to catch the prescribing error despite passing through 50 different steps and multiple checks before it was dispensed.

Image credit: Liu Tao, via Creative Commons

Prescribing errors carry a kind of momentum. The more checks an error slips through, the less likely it is to be doubted at subsequent checks. Similarly, diagnostic errors can exhibit *diagnostic momentum*. Once a diagnosis disseminates across the care team it becomes less likely to be questioned and harder to shake once found to be false.

Human-machine interaction: automation bias

Humans show similar behaviors when interacting with machines performing automated tasks, known as automation bias:

> Automation bias: "*The tendency to disregard or not search for contradictory information in light of a computer-generated solution that is accepted as correct" (Parasuraman & Riley, 1997)*

The study of automation bias in medicine has a rich history, but has become particularly relevant with the new machine learning approaches entering clinical decision support. The lack of transparency in *how* these models arrive at a decision will present a challenge for eliciting trust in clinicians and avoiding automation bias.

Automated systems exist on a spectrum from those that need human involvement, to full automation where humans are left out of any decision-making.

Full automation works for tasks that require no flexibility in decision making, have a low probability of failure and are low risk. But for dynamic environments where decision-making involves many changing variables — such as health care — full automation is much harder to achieve.

The performance of a system when applied to messy, real-world clinical practice will no doubt have an accuracy of less than 100%, so human operators must know when to trust, and when not to trust the system. How we design the interaction between the human and machine becomes of utmost importance to prevent *new* biases and errors being introduced.

Risk homeostasis would suggest that over-automating aspects of clinical practice might lead to complacency and an increase in errors and accidents. Studies from other fields have shown that humans do indeed suffer from automation bias and a reduction in personal accountability when their tasks are taken over by machines.

Acting in error vs. failing to act

Over-trust in an imperfect automated system leads to two specific types of errors: errors of *commission* and errors of *omission*. Errors of commission occur when a person acts erroneously, and errors of omission occur when the person fails to act when they should have.

A recent study investigated these two errors in a task using a decision support system. When the system provided the correct decision support recommendation, the participants' decision-making was faster, more accurate and required a lower cognitive load. But when the system gave the an incorrect recommendation ("automation wrong"), participant's decision-making performance dropped to near zero. The participants assumed the system was correct and made errors of *commission* — they *acted incorrectly.* When the system simply failed to give any recommendation at all ("automation gone") participants were more likely to make an error of *omission* — they *failed to act* when they should have.

Interestingly, the participant's accuracy dropped to a greater extent in the "automation wrong" condition compared to the "automation gone" condition. We make more errors when using a decision support system that is wrong (automation bias), compared to when a decision support system fails and we are left to our own devices (errors here are due to 'automation complacency'). The authors recommended that to avoid automation bias decision support systems should reveal to the user their lower degree of confidence when they are uncertain.

This kind of transparency will be essential in designing intelligent clinical decision support tools. Despite the hubris around machine learning approaches these technologies are still be subject to error and bias (e.g. data set shift) and this *must* be communicated to the clinician to avoid automation bias.

One particular challenge with automating health care is that errors often go unreported and their impact is hard to measure. Health outcomes from an episode of care are also only beginning to be tracked with the advent of value based healthcare. This means that once incorporated into clinical practice, both the positive and potential negative impacts of intelligent decision support systems may be difficult to identify. This makes it all the more important that we get the design right upfront.