Looking for explanations

written by Jeremy Huggett | 02/08/2016

(US Food and Drug Administration – Public Domain)

In 2014 the European Union determined that a person's 'right to be forgotten' by Google's search was a basic human right, but it remains the subject of dispute. If requested, Google currently removes links to an individual's specific search result on any Google domain that is accessed from within Europe and on any European Google domain from wherever it is accessed. Google is currently appealing against a proposed extension to this which would require the right to be forgotten to be extended to searches across all Google domains regardless of location, so that something which might be perfectly legal in one country would be removed from sight because of the laws of another. Not surprisingly, Google sees this as a fundamental challenge to accessibility of information.

As if the 'right to be forgotten' was not problematic enough, the EU has recently published its General Data Protection Regulation 2016/679 to be introduced from 2018 which places limits on the use of automated processing for decisions taken concerning individuals and requires explanations to be provided where an adverse effect on an individual can be demonstrated (Goodman and Flaxman 2016). This seems like a good idea on the face of it – shouldn't a self-driving car be able to explain the circumstances behind a collision? Why wouldn't we want a computer system to explain its reasoning, whether it concerns access to credit or the acquisition of an insurance policy or the classification of an archaeological object?

The problem is, we might not like the answer – not because we disagree with it, but because we

don't/can't understand it. The idea that we might understand the reasoning followed by a computer in taking a decision or making a diagnosis is arguably rooted in an image of computer decision-making from the 1980s. Artificial intelligence in those days primarily consisted of variants of rule-based systems in which multiple rules of the 'if … then …' variety determined the traversal of a problem domain and which, working in reverse, could provide a logical and clear explanation for the conclusion reached. At the time, this fitted neatly alongside a processual approach to archaeological explanation and saw a short-lived flurry of interest in archaeological expert systems.

Such brute-force methods may have worked for relatively restricted problem areas, diagnostic or artefact classification systems in archaeology, for instance, and even those as complex as chess as demonstrated by IBM's Deep Blue victory over Gary Kasparov in 1997, but they rapidly ran out of steam as a problem's complexity grew exponentially. Unlike Deep Blue, for example, Google's DeepMind AlphaGo wasn't programmed to play Go – with more possible board configurations than there are atoms in the universe (some 10^{170}), it can't be solved by algorithms that search exhaustively for the best move. Instead, it learned using a general-purpose deep learning algorithm that allowed it to interpret the game's patterns, first using information from expert games, then refining its methods by playing itself across 50 computers (e.g. Metz 2016).

Unlike machine learning using rules and training examples, corrected and tweaked by the human experts, deep learning such as this is unsupervised, the system learning and creating its own abstractions without the need for direct human intervention. Its conceptualisation of the problem space is machine-based, and no more than an imitation of the human approach, so that were the system able to explain itself, that explanation would primarily consist of a mass of data rather than a justification as such. And yet Demis Hassibis, one of the founders of DeepMind and now Vice President of Google's AI projects, has great ambitions for such software:

> "The system could process much larger volumes of data and surface the structural insight to the human expert in a way that is much more efficient—or maybe not possible for the human expert … The system could even suggest a way forward that might point the human expert to a breakthrough." (quoted in Metz 2016).

To some, the creation of an incomprehensible computational black box is not an issue. For example, Yoshua Bengio (in Pearson 2016) recently drew a comparison between a human expert and a computer, arguing that we don't understand the human and yet we trust them, so why should we not extend the same degree of trust to an expert computer? The difference is perhaps simply that the human can (or should be able to!) explain their reasoning in human terms – a deep learning AI system cannot do so (at present, at least). Indeed, the EU's GDPR specifies the need for human intervention in the explanation of automated decisions. This is not an alarmist argument – the reason why this is important is that that these complex systems may draw the wrong conclusions from complex data, and not necessarily obviously so.

For instance, Google's 'brain simulation' was famously able to detect cats within images without having been provided any prior information about them using deep learning techniques (e.g. Clark 2012). The system was provided with large quantities of data and allowed to learn from it. Similar software has been extended to identify the content of entire scenes. However, last month it was reported that facial recognition algorithms, which had performed very successfully and become

trusted as a result, broke down when presented with a very large dataset of one million faces (Hsu 2016). No specific reason why this might be the case is presented (Kemelmacher-Shlizerman *et al.* 2016) other than problems identified with photos of the same person at different ages and issues with children. It may be that this is another example of the kinds of problems identified with big data I highlighted in a previous post. A different but related issue has recently arisen with the mapping of the human brain: at the same time as what was claimed as the most detailed map of the human cortex was published (Fan 2016) it was shown that the functional Magnetic Resonance Imaging (fMRI) software used in measuring and mapping brain activity had fundamental flaws in its application (Oxenham 2016).

What all this underlines is that there are dangers in drawing conclusions based on systems which rely on large quantities of data being dropped into them and the software allowed to sift, classify, reason, and generate 'knowledge' as an output. This is not to suggest that there is no place for automated systems as some interpretations of the provisions of the EU GDPR would seem to suggest – instead it is an argument for the systems to be capable of providing human-scale explanations in support of their conclusions. It is not enough for the output to look right or even just look interesting unless the means by which it has been arrived at are accessible. And while there may not yet be an archaeological DeepMind or Watson, nevertheless many of the techniques used in these systems filter down and reappear in the automated methods applied to classify data, and to extract and abstract information from textual sources, for instance. We would expect an archaeologist to explain their reasoning in arriving at a conclusion; why should we not expect the same of a computer system?

References

Clark, L. 2012 'Google brain simulator identifies cats on YouTube', *Wired* (26th June). http://www.wired.co.uk/article/google-brain-recognises-cats

Fan, S. 2016 'Scientists Complete the Most Detailed Map of the Brain Ever', *Singularity Hub* (31st July). http://singularityhub.com/2016/07/31/scientists-complete-the-most-detailed-map-of-the-brain-ever/

Goodman, B. and Flaxman, S. 2016 'European Union regulations on algorithmic decision-making and a "right to be forgotten"', *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY, USA. https://arxiv.org/pdf/1606.08813.pdf

Hsu, J. 2016 'One Million Faces Challenge Even the Best Facial Recognition Algorithms', *IEEE Spectrum* (1st July). http://spectrum.ieee.org/tech-talk/computing/software/one-million-faces-challenge-even-the-best-facial-recognition-algorithms

Kemelmacher-Shlizerman, I., Seitz, S., Miller, D. and Brossard,E. 2016 'The MegaFace Benchmark: 1 Million Faces for Recognition at Scale', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. http://megaface.cs.washington.edu/KemelmacherMegaFaceCVPR16.pdf

Metz, C. 2016 'In a huge breakthrough, Google's AI beats a top player at the game of Go', *Wired* (27th

January). http://www.wired.com/2016/01/in-a-huge-breakthrough-googles-ai-beats-a-top-player-at-the-game-of-go/

Oxenham, S. 2016 ‘Thousands of fMRI brain studies in doubt due to software flaw’, *New Scientist* (18th
July). https://www.newscientist.com/article/2097734-thousands-of-fmri-brain-studies-in-doubt-due-to-software-flaws/

Pearson, J. 2016 ‘When AI goes wrong, we won’t be able to ask it why’, *Motherboard* (6th July). http://motherboard.vice.com/en_uk/read/ai-deep-learning-ethics-right-to-explanation