We the Robots? Chapter 6: Transparency

Pg 145:

For human users, the ability to know whether they are interacting with another human is coming to be seen as a basic right. This is a threshold question of transparency and not particularly controversial. Larger questions of transparency relate to the increasing opacity of AI systems.

As chapter three argued, this poses discrete challenges since it may allow inferior, impermissible, and illegitimate decisions to be made without the kind of scrutiny and accountability that would accompany human conduct of a similar nature.

Pg 146:

For human users, the ability to know whether they are interacting with another human is coming to be seen as a basic right. This is a threshold question of transparency and not particularly controversial. Larger questions of transparency relate to the increasing opacity of AI systems.

As chapter three argued, this poses discrete challenges since it may allow inferior, impermissible, and illegitimate decisions to be made without the kind of scrutiny and accountability that would accompany human conduct of a similar nature.

Pg 146:

An explanation, in this context, means a description of how certain factors were used to reach a particular decision. In order to be useful, it must be comprehensible and enable an interested person to understand the extent to which specific inputs influenced the output. This includes what factors were used and whether changing one or more of them would have yielded a different result; it should also enable a comparison between decisions, revealing the reasons for the difference or similarity.

A reasonable aim, perhaps, but it presents two problems. The first is that it necessarily requires simplification of the original system to make it comprehensible. The second is that it presumes that the purpose of an explanation is to help one individual understand a single decision. As we will see, that is only part of what explainability can mean – and only a fraction of what transparency should.

Pg 147:

A second approach therefore turned to instance-based explanations, also termed local or subject-centric interpretability: understanding the factors influencing a particular decision.

Pg 150:

The second aspect concerns the knowledge or expertise that can be presumed on the part of a regulator or user. It is commonly said that information disclosed must be ‘interpretable’, for example, in the sense of being able to be understood by a human. But just any human? There is a significant difference between explaining machine learning processes to a computer scientist and explaining them to a lay person, but there is no agreed technical standard for comprehensibility by a human - even though that is precisely the point of explainability. To the extent that only computer scientists are able to understand the work of their peers, putting technical experts in charge of accountability further runs the risk of regulatory capture.

As discussed in chapter three, even if a blanket prohibition on opaque AI systems were possible, it is not called for. Apart from anything else, a ban would mean that we forgo the many benefits that AI offers. Yet requiring that AI systems be ‘transparent’ also constrains innovation or introduces inefficiencies. Companies may be unwilling to expose trade secrets or invest the time and effort to develop sophisticated algorithms if they fear that these will be disclosed to competitors. Limiting the permissible number of variables in a model may render it more interpretable, but at the price of diminished accuracy.

Pg 152:

But this is not the same as transparency. A 2017 report by the US Defense Advanced Research Projects Agency (DARPA) similarly stated that the aims of XAI are enabling human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners’. Understanding and trust are important, though the focus on individual users is made evident in measurements of explanation effectiveness such as ‘user satisfaction’.

Pg 158:

An EU right to Explanation?

Previously known as the Article 29 Working Party, its guidelines on implementation of the GDPR provide that ‘meaningful information’ need not include a complex explanation of the algorithm or disclosure of the full algorithm but should be ‘sufficiently comprehensive for the data subject to understand the reasons for the decision’.

Pg 161:

The breaches by Google included its failure to provide information concerning the use of personal data in providing targeted advertising on its Android devices, leaving users unable ‘to sufficiently understand the particular consequences of the processing for them.

Pg 165:

Other jurisdictions have considered ways to preserve or encourage transparency while taking advantage of AI, though most have remained in the realm of voluntary principles comparable to Singapore’s Model Framework.

Pg 166:

But transparency also builds trust. That is routinely acknowledged to be one of the major barriers to adoption and acceptance of new technologies in general and AI in particular.

Pg 167:

Calls for transparency on the part of AI systems often start from questionable assumptions about human decision-making - contrasting algorithmic processing, for example, with traditional decision-making, where human decision-makers can in principle articulate their rationale when queried, limited only by their desire and capacity to give an explanation, and the questioner’s capacity to understand it’.

The ‘in principle’ is doing a lot of work here, as the process by which humans actually make decisions is known to be inextricably tied to intuition, hunches, personal impressions - with a layer of after-the-fact ratiocination. When we require a human decisionmaker to give reasons, we do not ask them to undergo functional magnetic resonance imaging in order to understand the cognitive process by which a decision was actually reached.

Language does not always help here. When considering explanations of different phenomena, we think of volitional human behaviour in terms of reasons rather than causes. When explaining a human decision, it would be odd to present the cause of a particular choice. Though we might say that new shoes cause us to walk in a particular way, we would not say that their discounted price ‘caused’ us to buy them. In the physical world, the reverse is true: we would not normally speak of the reason a fire started, except perhaps as the prelude to an explanation about a cause The language of ‘reasons’ presumes a degree of subjectivity and rationality on the part of an actor: they belong to that actor in a way that causes do not.

Transparency in AI systems is sought not for its own sake but for purposes similar to why it is sought in human decisions. The methods of achieving it are distinct, however. Human explanations emphasize factors influencing a decision rather than raw probabilities and are expressed in a manner that is tailored to the world views of the parties concerned. None of this is easy for an AI system. And sometimes the difference between AI and human explanations can be misleading. Where there is discretion to be exercised, for example, it can be artificial to ascribe reasons.