FUTURE AGENDA

Open Foresight

FUTURE VALUE OF DATA

An initial perspective

This paper is intended to provide a provocation for discussion. It raises issues without necessarily attempting to resolve them, and at times is deliberately controversial. We hope that it provides a point of departure for a meaningful conversation between different stakeholders, about the role and value of data in society.

Why are we talking about data?

The advent of 'big' has changed our relationship with data. In particular, the meteoric rise of the so-called 'tech titans' whose business models rely on the collection, creation and monetisation of huge data sets, has thrust data to the forefront of social and political discourses around the world. These companies, whose products are now woven into the very fabric of our existence, have shown us what data can do and how it can transform our lives, but perhaps unwittingly, they have also pushed a topic once the preserve of 'nerds' and 'wonks' into the mainstream. Global public debates around everything from growing inequalities, to political freedoms and human rights, and the very future of economic and social progress, all now involve heady proclamations about the use, abuse, power and possibility of big data.

To be clear, most of the current rhetoric in which data now finds itself playing a starring role is not new. Since William the Conqueror surveyed the lands, landholders and assets of England in 1085, and collected the data into volumes that came to be known as the Domesday Book ("the book of judgment"), the role of data in governance has been understood and argued over. A glance at the provenance of the German state's tough privacy protection laws also quickly reveals that the troubled dynamic between personal data collection and personal liberty was an issue that arose in a pre-

digital era. Scientific breakthroughs and social advancements such as space flight, or the discovery of new medicines and weather prediction, have all also given us a clear sense of the relationship between data and social progress, again, all long before the era of 'big' data.

What has changed, is the scale. Where once datasets were deliberate, discrete and focussed; 'big' datasets are huge, unwieldy, often indiscriminate, and of course, growing continuously and exponentially. What's more, we are now all in them. We feed them, knowingly or unknowingly, willingly or unwillingly, all the time. What has made this possible, is the relative ease and nonchalance with which the digital data trails of personal, social and physical processes can be collected and stored for analysis, now or at any time in the future.

It particular, the mass collection of 'personal' data that has brought forth a new kind of politics. It is the movement of data collection and analysis, experiment and discovery from remote and singular processes, to the most intimate and fundamental parts of everyone's personal, social and economic lives, seemingly without limit and without end, that has driven the idea of data into the heart of contemporary social and political discourse.

Defining data

One feature of much of the current debate around data is that it rarely seems to question what data actually is. Discussions and debates are held in which it is generally just assumed that everyone agrees and understands what it is they are discussing. But what exactly is data ?

The question is surprisingly difficult to answer, especially given how frequently the term is used. In fact, when using the term, we may actually be presuming a shared understanding between stakeholders that may not actually exist. When we use the term 'data', we need to think about whether we are all answering certain questions in the same way. For example, can data be owned? Is it an asset? Is it a raw material? Is it a product? Or is it a tool? Or more philosophically, does it only exist when it is captured, or does data have an independent existence outside of a data set?

Dictionary definitions of data might seem simple (that data is just 'information') but they are often not helpful. On the one hand it simply shifts the question to what exactly information is, and on the other, it is no more explanatory of the role of data in society, than would be a description of Shakespeare's Hamlet as just 'a collection of words'.

Even some of the words used to describe it that can seem unproblematic also fall foul of data's seemingly entirely context dependent nature. For example, many governments and organisations might see the description of data as an 'asset' as being fairly unproblematic. After all, data can be used to build new products, deliver better services, improve efficiency etc. According to this view, the more of it an organisation has, and the more it can protect and keep it for itself, the better it can innovate and create value or compete in a data-driven economy and polity. It is therefore an asset.

A short conversation with a data security expert might quickly shift this view however. Data security experts might argue that a vast amount of the data organisations hold should be better thought of as a liability, since the value they can extract from it is minimal in comparison to the costs of preventing it from being stolen or misused, or paying the price when it eventually, and inevitably, is. To give a concrete example of this, one might consider the situation in which organisations keep data related to online accounts that their customers have chosen to delete, or keep hold of data that has gone out of date (such as previous names, addresses, credit card details etc.). This data may be of some sort of use to that organisation at some point in the future, but will it prove to be more valuable than the risks incurred both legal and reputational, by storing it indefinitely? And what of it's potential to be misused by criminals or adversaries?

In the absence of good definitions of what exactly data is, or even a simple language that can adequately capture the myriad roles of data in modern society, cod analogies seem to have become commonplace. We will all be familiar, for example, with definitive-sounding statements such as 'data is the new oil' or 'data is the new currency'. These analogies can be useful. They can capture certain aspects of the way that data behaves in relation to certain aspects of economy and society, at certain times, but they equally mislead, causing us to associate data with things that it is decidedly unlike.

Is data like oil? Well, data is mined and refined, like oil. Vast hordes of it can make its owners (or 'controllers') very wealthy and powerful, like oil. We might even go to war over it, like oil. But there are also many ways in which data is not like oil. Data is not a finite, exhaustible resource, unlike oil. In many cases data is replicable or reproducible, unlike oil. The material costs of extraction, collection and movement of data are not high, unlike oil. The risks of data collection and use to society are real but not inherent, as they are with oil. Data ownership is not easily defined, unlike oil. Etc.

These differences are important since they point to a completely different set of end-points for the data economy than there have been for the oil economy, and so demand a different set of societal responses. The metaphor blinds us, in fact, to the different options we have around how we, as a society, might benefit from big data and avoid the calamitous potentials of it's use, in ways that are simply not possible when it comes to oil.

Is data like currency? It can certainly serve as a medium for exchange, as it does when a consumer, for example, shares their personal data in exchange for so-called 'free' services. It can also be used as a store of value, even in a quite a literal (albeit unstable) sense when it comes to crypto-currencies. So yes, data is like currency. But describing data as currency really doesn't tell us much. It just tells us that data has exchangeable value in certain contexts. In that sense, many things operate like currency. The economic value of data might have risen in recent times, and more people might be aware of that value, but the same might also be said of quinoa. Describing data as currency simply edits out many of its most important features.

During initial Future Agenda discussions we have heard other data-analogues: water (abundant and essential), sap (nourishing, flowing and replenishable), nuclear-weapons (dormant, but with the potential to cause catastrophe), life-support (as more and more essential systems and services become data reliant), perpetual memory (since data tends to bring the past into the present) and so on. All are worth exploring, as they each help to elucidate different aspects of the way data behaves in different contexts.

Perhaps searching for a single definition of data is a pointless exercise? The question of what data is may always depend on context. After all, even when using the term 'data' we are already eliding numerous different types of data and data-set (raw, cleaned, aggregated, anonymised, qualitative, quantitative, structured, unstructured, static, personal, processed, meta, open, closed, shared etc.) that data-analysts, statisticians and scientists have always understood to be quite different things.

If we are to accept this however, then we must also recognise the challenge of trying to take part in conversations and debates around data that involve multiple different perspectives, and in which different participants and stakeholders may hold very different ideas and knowledge-sets to those we hold. Taking a dogmatic approach and insisting that one has a more privileged view or understanding of data than others is only likely to lead to ever more extreme views as different stakeholders entrench on different sides of an argument.
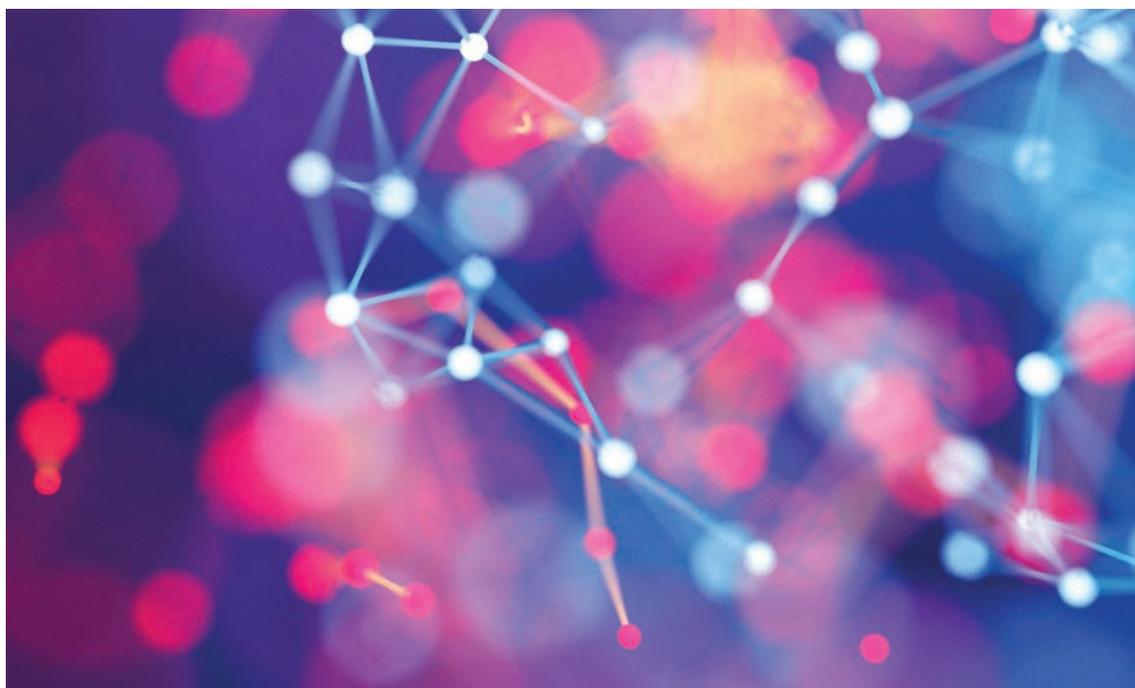
The current state of play

The landscape is polarised. Discussions around the role of data in society are often antagonistic, casting data's various roles and capabilities as either inherently good, or inherently bad, leaving little room for a middle ground.

Take for example, the heated politics surrounding encryption and personal messaging services. In essence this is a debate that sets the value of personal data privacy against the value of personal data collection to national security. For some, a regulated erosion of privacy is a necessary and reasonable price to pay for heightened national security; for others, even small erosions of privacy set us on the slippery slope towards the kind of society that is to be feared far more than any piecemeal threats to national security, and whose most unpleasant features can already be seen emerging around the world.

Another example might be the debate around data ownership. Much of the tenor of this debate, particularly around the commercial, governmental or entrepreneurial use of harvested personal data, is couched in a language that suggests that either 'you own it' or 'they own it', and never the twain shall meet. The reality is much messier than such rhetoric allows. When it comes to data and the law, for example, trying to sort out who exactly owns what exactly, and what that 'ownership' might entail, is no easy task (regardless of how forcefully the arguments over who should own data are made).

Similarly, whilst some argue that the principle of open data (particularly open government data) offers the best chance of unlocking the potential of big-data to solve societal challenges and bring collective benefit, others describe the exact same effort as giving away our most valuable assets to those with the best means to exploit it, whether or not they have the means to properly determine the best outcomes for society. The recent and controversial collaboration between the UK's National Health Service and Google's Deepmind division is a case in point. The partnership seemed to point towards exactly the kinds of optimistic hopes for big data sets and machine learning to help solve collective problems, whilst simultaneously sparking all of the worries around the potential harms of big data sets of personal information being collected and used by powerful stakeholders with inscrutable long-term interests.

The problem with debates around data taking on such antagonistic structures is that we could find ourselves in a position where public policies and data-strategies shift wildly between extremes depending on the differing ideological standpoints taken in different jurisdictions and markets, or within single jurisdictions over time, as policy and decision makers respond to fickle, event-driven public opinion.

What is the "value of data"?

It is clear that data can be used to drive both social and economic value. And, without getting lost in a metaphysical discussion about the concept of value, it seems safe to say that therefore the value of data lies in the uses to which it is put. Some of those uses seem to provide unequivocally positive value, such as the use of large data sets to build smart energy or water grids, to improve travel safety, or to search for new cures for diseases. Similarly, there are some uses of data which seem to generate unequivocally negative value like identity theft, cyber-attack, data blackmail, or the proliferation of false information ("fake news"). Other uses seem to allow for the generation of both positive and negative value, at the same time. Connecting people at a massive scale, for example, can enhance human relationships, allow ideas to flourish and give voice to those who may not otherwise have one; but it can also enable bullying, criminality or terrorism, give strength and credibility to bad ideas or ideologies and encourage mob rule. And there is also a category of data use which seems to create value that can be either positive or negative, depending on one's point of view. Data- harvesting for surveillance purposes, for example, or to develop new kinds of consumer products, or for the purposes of delivering targeted advertising, or for feeding sophisticated algorithms that underlie the efficient delivery of services (policing, insurance, access to government services etc.).

It should also be forcefully stated that the value (positive and negative) that data and it's uses can create, is not trivial. The world's wealthiest and most valuable companies are almost all now data-driven, or data-rich and the future of government looks set to be defined by 'smart' uses of large data sets. Consumers and citizens are beginning to understand this. Increasingly they are grasping the fact that what they once thought of as inconsequential personal data points, are actually being used to shape and define their lives at the very largest scales and are increasingly seeking ways to derive their own value from them.

It is no wonder then that we are currently bearing witness to a kind of data land grab in which companies, organisations, governments are all seeking to capture, horde and protect huge amounts of data from across the spectrum of human activity, with consumers and citizens perhaps only just beginning to resist and try to take control of their data for themselves.

In this context we are left with some important and urgent questions. What types of data, and what contexts, can be used to drive the most positive value from data? To whom is that positive value accruing? Who is best placed to use data to drive positive social value? What are the trade-offs and downsides of mass data collection, storage and use, and critically, who is monitoring or accountable for these?

In order to answer these questions it may be necessary to specifically explore the data value eco-system and explicitly call out the positive and negative values of data capture and use. For example, what exactly are the positive benefits of the collection of personal data, and what are the potential privacy harms that offset those benefits? What exactly is the social value that has been realised using big data sets, and what have been the social harms that have come about using the same? Only with honest reflection and evaluation can we truly consider whether the potential positive value outweighs the risks of potential negative value.

Where does this leave us?

There seem to be two questions that need to be urgently addressed. The first is to explore what a common language or framework for understanding data might be, so as to enable productive discussion between different data stakeholders. The second is to address the question of value in relation to the standard data questions: How is data created? Collected? Stored? Refined? By whom? And for what purposes?

One possible course of action might be to go back to first principles and explore the kinds of things that we can confidently say about data: That data's value seems to lie in the uses to which it is put (rather than, say, it's scarcity); That data can also be useless and dangerous; That data is abundant; That data can be stored indefinitely; That data is stored in physical locations, Etc. Perhaps by doing this we can build a set of common understandings on which other arguments can be layered.

The future

Data and the uses to which it is put are set to define not only the near term future for societies and economies but also the long-term future of humanity. We are going to see more data-driven companies, more cyber- security incidents, more breaches of privacy, more artificial intelligences, more miraculous transformations of the ways we live and more dramatic consequences of that transformation. Debates around data are not going to go away. They are going to intensify.

In the short term, properly or improperly, many of the mysteries around data and its role in societal and economic change are going to disappear. Citizens, service users, consumers… people… are going to find a way to understand the value of their data to different organisations, and the different uses to which their data is put. This will happen regardless of debates about whether the way they understand it is technically correct or incorrect. This de-mystification is sometimes portrayed as a shift in power to the consumer, but it is really about a simple conveyance of understanding of big data from the few to the many, and it will happen regardless of where power or wealth ultimately come to rest.

As this shift happens, those who can best grasp data's multiple possibilities and realities, it's multiple roles in, and value to, society, and render these things comprehensible to others, will likely have the more powerful voice.

The challenge then, is to solve some of the dilemmas we have raised in this paper. How might we approach the problem of bringing together different stakeholders with different perspectives to hold meaningful and constructive discussion? How can we develop a shared framework of understanding around the value of data? And when we talk about data in these diverse forums, how can we be sure we are talking about the same thing? We know that data is not simply 'the new oil'. So what is it?

FUTURE
AGENDA
Open Foresight