4.9 Data Quality

As we seek better insight, concern about biased, poor, and false data grows. Cleaning and validating data is a social, political, and commercial battleground.

Context

Whether it is basic administration, generating of new insights, making decisions, or organising their implementation, if the data that informs these activities is wrong, the outcome will almost certainly be sub-standard, inefficient, and potentially positively harmful.[129]

The most valuable data must be of good quality. Organisations clearly don't want bad quality data. Organisations that are in complete control of how their data is captured, indexed, and stored are in a better position to ensure quality, but for those that are seeking to combine information from external sources of varied quality and consistency, life can get tricky. That's why 'cleaning' data is big business. The question our workshops wanted to know is this: are we really rising to the challenge of poor data quality? If it is 'dirty', then all sorts of automated policies, investment, and even social decision-making may go astray; think of misaligned government funding due to inaccurate census data , children being wrongly removed from their parents because of an error in social service algorithms, or more mundanely, the duping of users on dating apps.

The Challenge

Our workshops distinguished between three types of low-quality data: poor, biased, and false.

- **Poor data** is incomplete, out of date, misattributed, misprocessed, or simply wrong. There are multiple reasons for this – from data entered into the wrong columns, to duplicate data or inconsistent entry, misspellings, and so on.

- **Biased data** refers more to sets of data that create a picture of something. This is now highly topical, as machine learning algorithms rely on these data sets to generate predictions and make decisions. A biased data set may simply reflect biases that already exist in society, such as the fact that most top jobs are held by middle-aged white men. But it can also reflect the values of coders, results from survey questions that are constructed with a particular slant, or arise from process/design issues such as data that is misreported in categorical groupings, non-random selections when sampling, or systematic measurement errors.

"We should focus on algorithmic awareness – NOT the elimination of bias, because we need to know why data was created."

Toronto workshop

Data Quality: Key Dimensions

- **False data** is deliberately created to be inaccurate or misleading - although it may well seem to be high quality and from verified sources. This has also become highly topical when false information is deliberately shared on social media, but it's also generated when individuals deliberately input false data because they don't trust the organisations that they are sharing the data with.

All three of these are now escalating in both scale and impact. They can render some data sets hard or impossible to use, and if not identified, corrected, and isolated, they end up polluting good data sets and the decisions based on them.

Managing Poor Data

Clarifying whether or not information is accurate is as yet largely a human, lengthy, and expensive task, although AI and wider automation is beginning to help. It explains why, in 2018, the global pharmaceutical company Roche was prepared to pay $1.9bn for Flatiron Health, a start-up which can clean clinical information with a particular focus on cancer. The capability that Roche valued in particular here, was the 'human-mediated extraction.'[130]

Many companies are grappling with how best to achieve better quality data, quickly, and at low cost. Some are focusing on improving data capture, and others are looking at ways to correct the errors. One option is only to use the good data and remove the 'bad' - but within this, it is important to define what 'good data' is. From a health perspective, for example, there is an emergent perspective that just because data is not of medical quality, does not mean it has no value. It's a question of what information is appropriate. This is a time-consuming and expensive exercise - 80% of data scientists' time is spent cleaning data.[131]

Biased Data

Most concerns about biased data focus on the data sets used to train and refine automated algorithms. In Washington DC, the case of an Amazon recruitment programme was discussed. Amazon's computer models were trained to vet applicants, by observing patterns in resumés submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry. The result was that the self-learning system taught itself that male candidates were preferable. There is no guarantee that other ways of sorting candidates that could prove discriminatory might occur – indeed, the Amazon algorithms allegedly also favoured men who played lacrosse and were called Jared.[132],[133] Amazon has since scrapped the project, but it's a good example of how difficult bias is to manage. Considering the fact that around 55 percent of US human resources managers expect to use AI within the next five years, this is extremely concerning in just this limited arena of recruitment.[134]

> There is a feedback loop – false data leads to low trust leads to false data."
>
> Hong Kong workshop

Another example discussed by the workshops was the claim that the AI algorithms currently used to decide who goes to jail are getting it wrong, due to their dependence on historical data.[135] In 2016, courtrooms in the US adopted risk assessment tools to generate a "recidivism score." This is decided by machine learning algorithms which use historical data to pick out the patterns associated with crime, to produce a single number estimating the likelihood of a prisoner reoffending. A judge then factors this into a prisoner's rehabilitation, or the duration of their sentence. This means populations that have historically been targeted by law enforcement, such as low-income and minority communities, are at risk of being given high recidivism scores. In turn, this means the algorithm could amplify embedded biases and generate even more bias to continue the cycle. Because most risk assessment algorithms are proprietary, it's also impossible to interrogate their decisions or hold them accountable.

Some in our workshops worried about a lack of diversity in the technology industry, and how this is impacting the roll-out of AI. Only 22% of AI professionals globally are female, for example. The more algorithms determine social outcomes, the more software development teams need to ensure diversity, to spot when data biases are skewing the decisions. Although there are increasing calls for more female coders, inventors, and investors, so that technology companies can more accurately reflect society, change is taking some time to come into effect. Some suspect there is a negative network effect, that the small share of women in the field discourages others from choosing it as a course of study. Employers might not be able to undo societies' gender bias single-handedly, but they can take mitigating steps, for example, by building tech skills into schemes for women returning from career breaks, and providing greater transparency around pay and opportunity.

AI can help expose truth inside messy data sets, and will be used to great benefit in multiple different ways. But it poses potential risks as well as opportunities. A frequent topic of conversation in our workshops was the need for business leaders to establish a transparent process for monitoring the ethical behaviour of their AI systems. This could include common standards for training data for algorithm building and real-world applications. Part of the solution may also lie in regulation, including hefty fines for non-compliance, plus a concerted effort to ensure that there is greater public awareness of the potential issues.

"Labelling helps to identify truth, and perhaps branded news is a way to help the public identify responsible channels."

Mexico City workshop

False Data

'Fake news' is now big news, and a major headache for both tech companies and governments. There is a large and growing market for exploiting the vulnerabilities of the digital world, and some very smart, sometimes unscrupulous, players capable of supplying it. Such is the sophistication of some of the false information, that it can be almost impossible to identify it. Campaigners are pushing governments to develop tougher regulation to better protect civil society. Some are considering adopting tighter international protocols, such as those used to restrict the arms trade.

Much of this debate is beyond the scope of this report, but fake news is not the only form of false data. In our Washington DC discussion, for example, it was pointed out that around 20% of US Census data is thought to be inaccurate, mostly because citizens providing the information fear how government will react if they tell the truth. Officials for the US Census are not allowed to compensate for this, despite knowing that around 20% of the key data sets are wrong. Here, the inaccurate data is largely driven by public fear of government intervention, and some communities; often those in most need of support, such as the poor, recent immigrants, and the elderly, intentionally enter false data for personal information like income, health, and age. The unfortunate irony is, without census data to identify need, policy makers are unable to justify additional funds to support the very people who are not disclosing the correct data. When we discussed this a few days later in Toronto, there was an acknowledgement of similar statistical issues, but officials in Canada are allowed to 'correct' known data sets before they lead to ineffective policy and misguided activity.

It may not matter much if we give false email addresses to access public wi-fi, or when shopping for a new pair of shoes, but it does when there are important consequences. In Nigeria, such is the level of mistrust, that few give government agencies accurate information or correct emails. As was observed in Hong Kong, *"there is a feedback loop – false data leads to low trust leads to false data."* The challenge comes when data has to be real enough to authenticate an individual, a machine, or a location. The principle of digital identity is important here, and has recently been explored in detail in another Future Agenda project.[136]

We must be careful not to make the perfect the enemy of the good. Just because you identify bias, doesn't mean it is inherently flawed."

Santiago workshop

What We Heard

Across the world, there was deep concern about the provenance and accuracy of information served up to individuals on social media, and the role of algorithms in this. In Bangalore, there was a *"growing concern how to monitor and control social media, to limit the manipulation of consumers by corporates and other organisations."* In Mexico City, a concern was that *"discrimination will be a big issue – particularly as facial recognition becomes more prevalent."* Singapore was more optimistic, *"AI will become more sophisticated around helping identify fraudsters, but we are not sure if it will be fast enough to identify fake news before it gets out …. Labelling helps to identify truth, and perhaps branded news is a way to help the public identify responsible channels."*

Many mentioned the seemingly blind confidence that there is in the accuracy of algorithms, and observed that even clean data can be biased. In Madrid, several highlighted that *"biased data is increasingly powering automated choices."*[137] In Canada, the suggestion was that bias should be managed through *"algorithmic awareness – NOT the elimination of bias, because we need to know why data was created."*

In Santiago, it was suggested that *"we need to work out if it is at all possible to measure bias."* Is it possible to develop a quality mark or traffic lights system for data, showing whether or not it is free from bias, moderately impacted, or severely compromised? However, in Hong Kong, the view was that *"we must be careful not to make the perfect the enemy of the good. Just because you identify bias, doesn't mean it is inherently flawed."* That said, in the same workshop, it was acknowledged that *"there is a risk that bias will be programmed into AI, which will lead to continuous marginalisation of individuals."* What is certainly the case is that, given machine learning is retrospective, the more we rely on machine learning, the more existing bias can potentially be entrenched.

One suggested solution was to *"consider developing strong regulation frameworks that require harm-based assessments of the application of data, and continues to monitor real-world harm."* Some in Hong Kong also wondered, *"should there be a world data organisation that can establish principles around bias?"* There, it was also proposed that *"the key question is which institution will be able to identify and exclude bias, both of input and output. Do facts need to be baked into this?"* Additionally, "*it is difficult at this point to identify whether the outcome will be positive or negative. There are plenty of examples of bias in China, around many issues – from mortgages and AIDS, to sentencing, diversity, and inclusion - and it is difficult to see how individuals have been categorised."*

Another thought in Sydney was that *"bias within data could lead to data inequality."* Looking forward to 2030 in London, some agreed, and saw that we will see *"more social exclusion in terms of in-built bias of automated process, networks, and creators."*

> "The challenge will be to extend legal protection over all aspects of life; for example, the wide range of potential cases which may have a discriminatory outcome that affect people or third parties."
>
> Santiago workshop

In Nigeria, the problem is more societal, as *"corruption and lack of trust in the system is driving the collection of inaccurate and fake data."* People intentionally give false information to the government and companies alike. This *"makes our databases unreliable, as citizens choose not to share accurate information."* Other than eliminating corruption, suggestions of how to overcome this focused on better public education to *"build a wider understanding of the benefits of data sharing."*

Looking Forward

The assessment from Copenhagen was that *"at the core, we need to have objective views of what is good data - but being clear on what is this 'objectivity' is a central question… a big issue for the future is who will decide."* They acknowledged that *"for public consensus, we may have to go through a period of more data anarchy and more fake data, before people change."*

The final workshop in Santiago agreed, *"between today and 2030, existing regulation needs to be updated. Policy makers need to be trained on this and so be able to agree on the appropriate use of algorithms, and to better identify instances of bias as a start."* We also need to consider taxonomy and how we classify algorithms; *"the challenge will be to extend legal protection over all aspects of life; for example, the wide range of potential cases which may have a discriminatory outcome that affect people or third parties."*

Some argued for a *"World Data Organisation, which can establish principles about quality and bias."*[138] However, controlling the spread of fake data is more challenging. It contaminates good data sets, distorts our perspective, and gradually misleads our actions.

Implications for Data Value

If our data in the future is to be useable, never mind of value to society and commerce alike, then it has to be reliable. The view from those who discussed this in our workshops was that society hasn't yet acknowledged either the scale or the complexity of this problem. Improved transparency and accountability processes can help, but it is also about underlying data quality. However, acknowledging and managing raw and contaminated data alongside cleaned data, is a necessary shift that many will need to accommodate. For most requirements, some inaccuracies can be managed, but certainly not all – think of clinical trials results, for example. Global consensus around acceptable levels of accuracy would help here, alongside an institution which can set standards and then arbitrate should disagreement arise.

It is clear that organisations that can efficiently, quickly, and accurately clean data are already adding value, and that high quality, structured data sets will continue to command a premium. As data becomes even more integrated into the operations of our economy and society, it is increasingly important to ensure and maintain its quality.

"Corruption and lack of trust in the system is driving the collection of inaccurate and fake data."

Abuja workshop

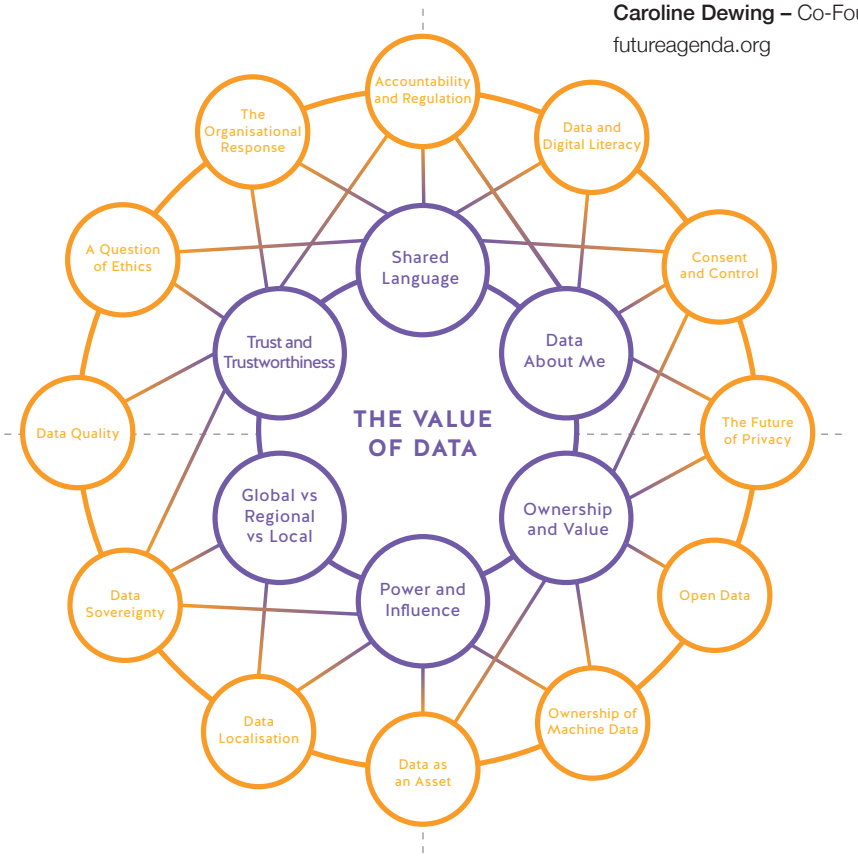
Context

This is one of 18 key insights to emerge from a major global open foresight project exploring the future value of data.

Throughout 2018, Future Agenda canvassed the views of a wide range of 900 experts with different backgrounds and perspectives from around the world, to provide their insights on the future value of data. Supported by Facebook and many other organisations, we held 30 workshops across 24 countries in Africa, Asia, the Americas, and Europe. In them, we reviewed the data landscape across the globe, as it is now, and how experts think it will evolve over the next five to ten years.

The aim of the project was to gain a better understanding of how perspectives and priorities differ across the world, and to use the diverse voices and viewpoints to help governments, organisations, and individuals to better understand what they need to do to realise data's full potential.

From the multiple discussions 6 over-arching themes were identified alongside 12 additional, related future shifts as summarised in the diagram below.



Details of each of these, a full report and additional supporting information can all be found on the dedicated mini-site: www.deliveringvaluethroughdata.org

About Future Agenda

Future Agenda is an open source think tank and advisory firm. It runs a global open foresight programme, helping organisations to identify emerging opportunities, and make more informed decisions. Future Agenda also supports leading organisations, large and small, on strategy, growth and innovation.

Founded in 2010, Future Agenda has pioneered an open foresight approach bringing together senior leaders across business, academia, NFP and government to challenge assumptions about the next ten years, build an informed view and establish robust growth strategies focused on major emerging opportunities. We connect the informed and influential to help drive lasting impact.

For more information please see:
www.futureagenda.org

For more details of this project contact:
Dr Tim Jones – Programme Director, tim.jones@futureagenda.org
Caroline Dewing – Co-Founder, caroline.dewing@futureagenda.org

Text © Future Agenda
Images © istockimages.com
First published November 2019 by:
Future Agenda Limited
84 Brook Street
London
W1K 5EH