The role of data in AI business models

Open Data Institute

About

This report has been researched and produced by the Open Data Institute, and published in April 2018. Its lead authors were Jared Robert Keller, Lucia Chauvet, Jamie Fawcett and Olivier Thereaux, with contributions from Caley Dewhurst, Anna Scott, Peter Wells and Jeni Tennison.

If you would like to send us feedback, please get in touch by email at RandD@theodi.org.

Contents

Executive summary

The current wave of interest and investment in artificial intelligence (AI) is characterised in part by a preference for business models that prioritise the collection and siloing of data. For companies taking this approach, access to data is seen as the key to securing a competitive advantage when building, implementing, and operating AI systems. Many of these companies therefore choose to restrict access to data in an effort to increase their competitive advantage.

The most frequently cited justifications for this approach are that the data siloed by companies is personal or proprietary – that by restricting access to data, companies are ensuring privacy and increasing profits.

However, this approach – and the business models based on it – arguably undercuts these goals. It undercuts the goal of privacy by allowing large companies to control data about people rather than building a framework that recognises the needs of people, communities, businesses and governments. Along similar lines, the current approach undercuts efforts to limit unintended encoded bias in AI systems since it makes it difficult to interrogate the data used to train those systems. Finally, the widespread practice of restricting access to data undercuts the goal of increasing profits, since an AI sector beset by monopolies and the widespread siloing of crucial data stifles innovation within the sector as a whole.

In this report we will present a matrix of AI business models to help readers understand the array of models currently used by businesses developing AI systems, and lay out the most commonly expressed reasons for the current trend of restricting access to data. We will then discuss challenges related to the widespread siloing of data, focusing on the dangers of unintended encoded bias and an AI oligopoly. We will close by identifying a number of emerging trends that are likely to impact the ways data is used in the AI sector over the coming years and explore ways of mitigating against the dangers of oligopoly and unintended encoded bias by opening and sharing data.

At the ODI we believe the ideal path forward will involve increasing the amount of sharing and opening of data for and by businesses operating AI systems, as well as the adoption of a wider variety of business models within the AI community. This will need to be done in a way that secures and safeguards the trust of people while promoting innovation across the sector.

Introduction

Artificial intelligence systems are a combination of clever statistical and mathematical techniques, an understanding of the world, and lots of data.

Artificial intelligence (AI) has lately garnered a lot of attention and excitement, not only from private and public sector organisations but the general public as well. National governments, public sector organisations, large corporations, SMEs and startups have all invested significant amounts of money in developing AI technologies such as machine learning, deep learning, natural language processing and image recognition which, according to the Centre of Public Impact (CPI), can usefully be described as "software that enhances and automates the knowledge-based work done by humans".[1] In this report we have referred to these different approaches as 'AI systems'.

The largest technology companies and established firms are particularly interested in developing and implementing AI tools and frameworks. In 2011, IBM invested $15 billion USD in Watson, its cognitive AI system for answering natural language questions and deriving meaning from photos, videos, text and speech.[2] In addition to the money being spent by individual tech companies, countries such as the USA, China, the UK and France are also investing money in building AI industries. In 2017, for example, the UK government stated its intention to invest at least £75 million into the field of AI in order to "secure the UK's leading position in the global AI market".[3]

Some of this interest in AI is focused on the long-term potential and impact of general purpose AI and superintelligence. However, much of it centres on applying AI technologies to help tackle specific problems. The potential impact of these targeted applications has been estimated by PricewaterhouseCoopers (PWC) as adding up to $15.7 trillion USD to global GDP by 2030, making AI the "biggest commercial opportunity in today's fast changing economy".[4] From financial services to digital marketing and healthcare, artificial intelligence has already begun to demonstrate its potential in a variety of sectors, saving time, money and lives. In 2016, Frost and Sullivan, found that the application of AI solutions in the healthcare sector has the potential to "improve outcomes by 30 to 40 percent, while cutting treatment costs by as much as 50 percent."[5]

At the ODI, we are exploring how some or all of this potential might be realised through the application of AI systems to specific challenges. In this report, we focus on the role

[1] Centre for Public Impact (2017), 'Destination Unknown: Exploring the impact of Artificial Intelligence on Government', https://publicimpact.blob.core.windows.net/production/2017/09/Destination-Unknown-AI-and-government.pdf

[2] Frank Palermo (2016), 'IBM Watson Points the Way to Our Cognitive Business Future', https://www.cmswire.com/information-management/ibm-watson-points-the-way-to-our-cognitive-business-future/

[3] Department of HM Treasury (2017), 'Autumn Budget 2017', https://www.gov.uk/government/publications/autumn-budget-2017-documents/autumn-budget-2017

[4] PricewaterhouseCoopers (2017), 'AI to drive GDP gains of $15.7 trillion with productivity, personalisation improvements', https://press.pwc.com/News-releases/ai-to-drive-gdp-gains-of--15.7-trillion-with-productivity--personalisation-improvements/s/3cc702e4-9cac-4a17-85b9-71769fba82a6

[5] Frost & Sullivan (2016), 'From $600m to $6bn, Artificial Intelligence Systems Poised for Dramatic Market Expansion in Healthcare', https://ww2.frost.com/news/press-releases/600-m-6-billion-artificial-intelligence-systems-poised-dramatic-market-expansion-healthcare

of data in building such systems, and how access to data, or a lack of access, might impact the creation of value for people, businesses and society.

A brief history of artificial intelligence

Despite all the recent attention and excitement, the idea of ‘artificial intelligence’ is not new – indeed, the creation of artificial intelligence has been the subject of serious academic and industrial research since the emergence of electronic computing in the early post-war period. The intervening decades have seen periodic waves of excitement about the potential of AI systems.

> “ *There have been at least three or four massive waves of enthusiasm for AI since the sixties through to now, and on each occasion the headlines have been the same; ‘the robots will take our jobs’ and ‘human expertise is not what it was’, and ‘this pinnacle of human intelligence has just been defeated by a machine so what does that mean about us?’ And associated with each of those peaks have been the troughs that have followed the kind of classic Gartner life cycle*
>
> ***– SIr Nigel Shadbolt***

In that time, our understanding of what is meant by ‘artificial intelligence’ has changed numerous times, often leading to confusion about what is and isn’t ‘artificial intelligence’. According to Michael Veale, a researcher at UCL focusing on the interplay between machine learning and data protection, “*AI has always been characterised by people moving the goalposts. People say, ‘Hey, we’ll have an AI system when we can beat this person at chess’, and then of course it does, and people move the goalposts.*” That is why we prefer the CPI’s description of AI as “software that enhances and automates the knowledge-based work done by humans”.[6] In fact, according to Nigel Shadbolt, Chairman and Co-Founder of the Open Data Institute, Professor of artificial intelligence and Principal of Jesus College, Oxford:

> “ *John McCarthy said that “today’s ‘AI’ is tomorrow’s computer science” – which is to say that what looks like a really unique method today, just becomes one of the many methods in your library of computer science techniques.*

There are also a number of misconceptions about what AI systems are and what they can do. Chiefly, there is a tendency to conflate the potential for superintelligent computer systems that can perform tasks in a manner equivalent or better than

[6] Centre for Public Impact (2017), ‘Destination Unknown: Exploring the impact of Artificial Intelligence on Government’, https://publicimpact.blob.core.windows.net/production/2017/09/Destination-Unknown-AI-and-government.pdf

humans, with the more conventional programmes that carry out specific tasks based on some ‘learned’ behaviour. Terming both of these as ‘artificial intelligence’ can be confusing and has led some to suggest the need to rethink the terminology. Indeed, Yann Lechelle from Snips – which provides localised voice recognition systems – stated:

> “
> *AI alone is misleading because it’s a grab bag, it’s loaded with century old literature from Isaac Asimov, it’s loaded with Hollywood’s rendering of AI gone awry, it’s loaded with our collective fantasy. I think we need different words for AI.*

Naming things is, famously, one of the hardest things in computer science.[7] The term ‘artificial intelligence’ itself gained popularity in the 1950s having won out over "complex information processing"– a term that arguably comes closer to explaining what these systems actually do.

What is an AI system?

AI systems are executable models that take some data input and produce specific recommendations, decisions or actions. They are a combination of statistical and mathematical techniques, an understanding of the world and data. There are many ways of creating these models: expert systems codify the knowledge of experts while machine learning systems are built using statistical techniques trained over data. Individually they are perhaps better explained in terms of tasks they perform – such as automated decision making, natural language processing and pattern recognition – as opposed to the many methods through which they are created – such as machine learning or deep neural networks.

The present day

While AI systems may not be new, our interviewees argued that there are three factors that mark out the current wave of excitement surrounding AI: the increased processing power of today’s computers, the availability of open source toolkits and the large amount of data now available.

The recent explosion in AI development has less to do with the development of new algorithms than the availability of high-performance hardware. Most of the techniques used today, from neural networks to linear regression, were theorised and developed decades – and in some cases centuries – ago. It was only recently that hardware made those techniques usable at scale: storage technology is now mature enough to store and shift vast amounts of training data; the development of GPUs for graphics and gaming applications have made massive parallelised computing significantly cheaper than when neural networks were invented. More recently, the development and release of AI-specialised hardware such as Google’s Tensor Processing Unit (TPU) or Intel’s Nervana – which not only increase the performance of AI algorithms, but are often

[7] David P. Karlton (2017), ‘Naming Things is Hard’, http://www.meerkat.com/2017/12/naming-things-hard/

optimised for a specific AI algorithm[8] – are a sign of the maturity of the software-hardware ecosystem.

In parallel, according to Michael Veale, “there’s more data now than there was before, or at least companies have more structured forms of data.” Indeed, a report from IBM Marketing Cloud estimated that 90% of the data that exists today was created in the last two years, with around “2.5 quintillion bytes of data”[9] produced each day from almost every sector of the economy.[10] Alongside the increasing amounts of data being generated is a trend for increasing access to data, either through public availability on the web or explicit data sharing and open data publishing.

> “ *Much of AI’s recent success has come on the back of electrical engineers building better and faster chips, better and bigger memory as well as the network effect of the web and technology to generate torrents of content and data and the human ingenuity to have committed all of that to machine-usable form.*
>
> ***– SIr Nigel Shadbolt***

Finally, the increasing availability of and access to open source tools and frameworks for creating AI systems also play a part in the current wave of excitement. Tensorflow, Torch and Spark are examples of open source software libraries which, along with powerful cloud computing providers such as AWS and Google Cloud, have made the creation of AI systems – especially during research and development – significantly easier.

The importance of data in AI systems

AI systems are a combination of statistical and mathematical techniques, an understanding of the world and data – lots of data. Large amounts of quality data are crucial not only as an input during the operation of AI systems, but also to train them in the first place. Without access to lots of high quality data, algorithms cannot learn. Even the most powerful AI techniques – with top of the line hardware are significantly less useful without access to quality data.

The more data that companies have access to, and the better the quality, the more accurate their AI systems can be. Research carried out for this study showed that access to clean, varied and quality data is required to create effective AI systems. Perhaps more importantly, poor quality data or data that is not sufficiently varied, results in misleading or erroneous outputs from AI systems. According to a study by the Boston Consulting Group (BCG) and Massachusetts Institute of Technology (MIT) Sloan Management Review, a clear discriminator between ‘leaders’ and ‘laggards’ in

[8] Joel Hruska (2017), ‘Google’s dedicated TensorFlow processor, or TPU, crushes Intel, Nvidia in inference workloads’, https://www.extremetech.com/computing/247199-googles-dedicated-tensorflow-processor-tpu-makes-hash-intel-nvidia-inference-workloads.

[9] Jack Loechner (2016), ‘90% Of Today's Data Created In Two Years’, https://www.mediapost.com/publications/article/291358/90-of-todays-data-created-in-two-years.html

[10] Richard Harris (2016), ‘More data will be created in 2017 than the previous 5,000 years of humanity’, https://appdevelopermagazine.com/4773/2016/12/23/more-data-will-be-created-in-2017-than-the-previous-5,000-years-of-humanity-/

the application of AI systems to business processes is “their understanding of the importance of data, training, and algorithms”.[11] Effective, commercially viable, AI systems require:

> i) an understanding of the importance of data, and in particular an awareness of the sensitivity of algorithms to poor quality data, and
>
> ii) the creation of systems to integrate new data into the model.

According to a few of our interviewees, there is a common misconception that the most important part of an AI system is the algorithm, when in fact the data upon which the algorithm is trained and operated is just as important. One interviewee, who did not wish to be cited, believed the perpetuation of this misconception is due to the difficulty of explaining the role of data in both training and operating AI systems, and that focusing on the algorithm is “just an easier story to tell”.

Many of our interviewees echoed the importance of access to data. Stephen Whitworth, founder of the fraud detection startup, Ravelin, explained that if another business set out to develop an AI system that could rival Ravelin’s, they would do so not only by building a “slightly better algorithm”, but by gaining access to “better data”. John Enevoldsen from Ocean Protocol similarly noted that while having a good algorithm was a necessary and important part of any business developing an AI system, “the amount of relevant data you have behind it” is as important as the algorithm itself.

While almost all of our interviewees highlighted the importance of large quantities of data, nearly as many also emphasised that the data being used has to be appropriate – it has to fit the problem being tackled.

> *The value in most data is just in the massive quantity of it and how it represents the problem that you’re facing. The answer is in there somewhere. The neural network is simply a tool to go and try to find the answers, it’s not going to automatically solve anything for you*
>
> ***– Steve, minds.ai***

Because of the significant role of data in making effective AI systems, we will aim to explore the role access to data has in AI business models. This includes examining the current trends in access to data, identifying emerging challenges with current approaches and considering the potential effects of regulation and trust. Finally, we will look at how data sharing and opening of data might enable the emergence of a greater number of more effective AI systems.

[11] Philipp Gerbert et al. (2017), ‘Is Your Business Ready for Artificial Intelligence?’, https://www.bcg.com/publications/2017/strategy-technology-digital-is-your-business-ready-artificial-intelligence.aspx

Understanding AI business models

There are many strategies for creating a competitive advantage when building AI systems, but the two most prominent are controlling access to data and algorithms.

AI businesses can be found operating across a wide range of industries and sectors. For this research we spoke to 17 organisations working in industries as diverse as online credit fraud prevention, child protection, and transport, as well as companies who work across a number of commercial contexts simultaneously. We also spoke to companies that provide services to businesses deploying AI systems, including providing access to data, and venture capitalists who fund AI startups.

Across these sectors, there is a wide range of different products and services based on AI systems trying to meet a wide variety of different user needs. As a result there are many potential business models and techniques for creating competitive advantages in use, such as building relationships with customers and better meeting their needs. However, there are a set of similar characteristics that all business models around AI systems have. By focusing on these, we aim to understand how businesses are attempting to create a competitive advantage for their products and services through AI systems.

To create effective AI systems that can be deployed within various products and services, businesses need access to **algorithms, data, skills and hardware**. To create a competitive advantage in providing these products and services, companies often choose to restrict access to some or all of these resources. Because of the ubiquity and commodification of hardware, access to sufficient hardware is relatively uncompetitive, though it may have an impact on initial development done by startups.[12] Any business attempting to create a competitive advantage through the development of bespoke hardware would encounter large upfront costs, and in many applications this would be unlikely to provide a sufficient competitive advantage to justify that cost. However it is a strategy that can be employed by large corporations running cutting-edge AI systems.

While there is a well-documented scarcity of advanced AI skills and therefore extensive competition for AI expertise,[13] investment in recruiting top AI talent to create a competitive advantage is one potential strategy. In many cases, however, it might be prohibitively expensive for businesses developing AI systems, given the need for significant capital and potentially limited returns if the AI systems being developed are not cutting edge.

Creating a competitive advantage through access to data and algorithms

Because of the difficulty with building a competitive advantage through access to hardware and skills, access to data and access to algorithms are the primary means of creating a competitive advantage for most businesses creating AI systems.

[12] Digital Catapult Centre (2018), 'Machines for Machine Learning', https://www.digitalcatapultcentre.org.uk/wp-content/uploads/2018/01/MIG_MachinesforMachineIntelligence_Report_DigitalCatapult-1.pdf
[13] Rita C. Waite (2017), 'This Is Why All Companies Need An AI Strategy Today', https://www.cbinsights.com/research/artificial-intelligence-strategy

In traditional business best practice, businesses lock down access to their resources – however with access to data and algorithms this is not necessarily the desired approach. This is because both data and algorithms are digital assets – in economic terms they are 'non-rivalrous goods', meaning they can be used by many people without being used up.[14] Providing access to data or algorithms does not hinder a company's ability to use it themselves, so the perceived cost of sharing these assets is primarily attributed to a loss of advantage over competitors.

There are potential business benefits of sharing access to both data and algorithms – often centred on the potential for AI systems that are faster and of higher quality, as well as spurring greater innovation within the AI sector as a whole. Businesses may choose to enable access to data or algorithms in order to realise some of these benefits, however they also have to consider the impact this may have on their ability to retain a competitive advantage. In order to realise the potential benefits of wider access while retaining some form of competitive advantage, every business must make informed decisions about whether – and how widely – to share access to both the data they use and algorithms they create.

This challenge was echoed by one of our interviewees, Roemer Claasen of Frosha (a Dutch company using AI to clean and combine data) who said he and his colleagues often feel a tension between needing to protect their competitive advantage while at the same time wanting to realise the potential benefits of making their algorithms available:

> “ *We do think about making the algorithm usable, at least for the open data community but we have to be very careful because it's also our business model...*
>
> ***– Roemer Claasen, founder of Frosha***

Access to data and algorithms exist on a spectrum

Providing access to both data and algorithms is not a binary choice. Businesses need to decide what to share, with whom and how.

For data, access can be described by the ODI Data Spectrum – where access to data exists on a spectrum based on the terms of a data licence, sharing agreement or permissions. This spectrum spans from not sharing access with any external organisations to publishing as open data – with a licence that allows anyone to access, use and share. In between, there is a wide range of possible types of sharing – including directly with selected partners, with groups of partners or publicly with a restrictive licence.

Within an individual AI system, different data might be treated differently and therefore exist at different points on the spectrum – for example data used to train an algorithm might be closed because it contains personal data, but a subset of suitably anonymised data may be published openly. Examples of where data is being closed, shared or opened include:

- Frosha, which deals with highly sensitive data about people and keeps that data closed[15]

[14] University of Pittsburgh, 'Public Goods', http://www2.pitt.edu/~upjecon/MCG/MICRO/GOVT/Pubgood.html
[15] Frosha (2017), https://frosha.io

- Ravelin, which, as a service for clients, shares data which has been anonymised by its AI system[16]
- Mozilla, a not-for-profit technology company, that publishes its voice data set openly[17]

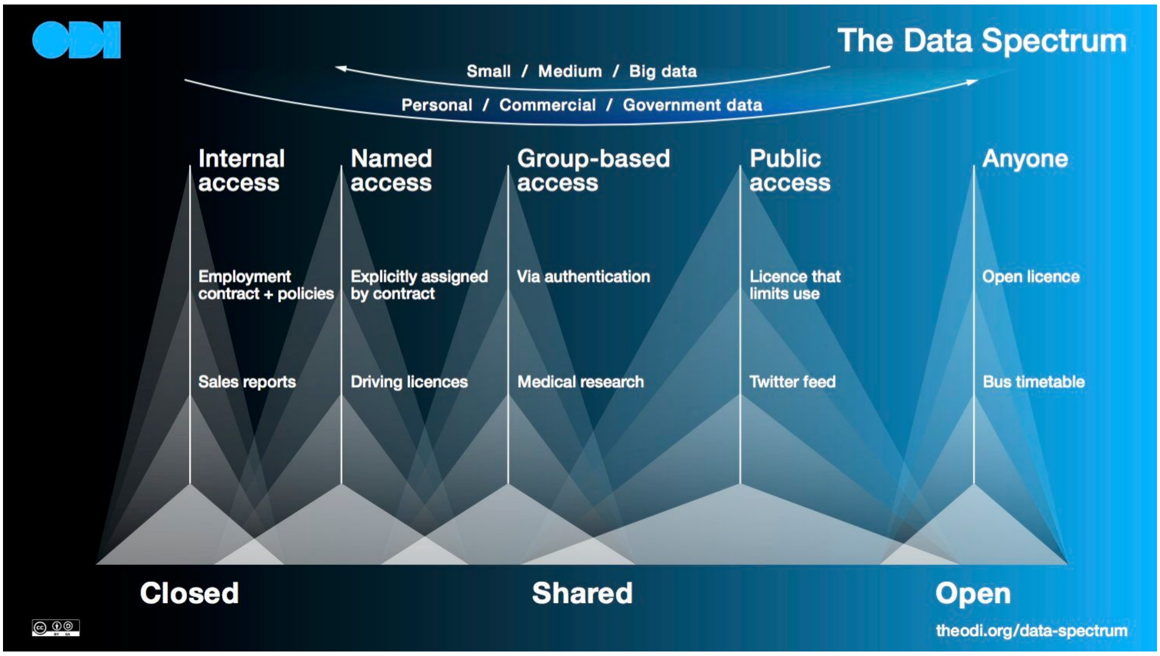

Figure 1. The Data Spectrum.

From our research, we concluded that access to algorithms follows a similar pattern as access to data. Algorithms can therefore be understood to exist on a spectrum from closed to shared to fully open. However, this spectrum is more complicated because the term ‘AI algorithm’ can be used to describe a range of different facets of an AI system, including the type of AI system (e.g. recurrent neural network, linear regression or decision trees, to name only a few popular examples), the software toolkit used to implement the system, the model created, the features selected, or indeed any combination of these. An individual ‘AI algorithm’ might, therefore, cover an area of the spectrum – for example, a business might publish the open source software framework used to create a specific AI model, but keep details of the feature definition or optimisation techniques closed.

While the complexity makes it more difficult to classify AI algorithms as existing at particular points on a spectrum of access, understanding the broad area where they exist enables comparisons between AI systems in terms of their business model and approach to competitive advantage.

For example, Snips keeps their algorithms closed because their commercial advantage relies on controlling access to their algorithms. A number of large technology providers share their algorithms by offering paid or restricted access to their AI systems, including Amazon Web Services which provides access to Lex, a service that allows users to build conversational interfaces using the algorithm that powers Amazon’s own virtual assistant, Alexa.[18] Finally, many of the larger technology companies contribute to

[16] Ravelin (2018), https://www.ravelin.com/ravelinlookup

[17] Sean White (2017), ‘Announcing the Initial Release of Mozilla’s Open Source Speech Recognition Model and Voice Dataset’, https://blog.mozilla.org/blog/2017/11/29/announcing-the-initial-release-of-mozillas-open-source-speech-recognition-model-and-voice-dataset

[18] Amazon Web Services (2018), ‘What is Amazon Lex?’, https://docs.aws.amazon.com/lex/latest/dg/what-is.html

open source AI toolkits, for example Microsoft's Computational Network Toolkit[19] or Google's TensorFlow.[20]

Classifying AI systems by access to data and algorithms

By considering products and services built on AI systems in terms of their position on spectrums of access to data and algorithms, we can start to understand how these play into their business model. If we overlay these spectrums in a matrix, we can, at least theoretically, place each product or service at a point on that matrix – or at least across an area of the matrix. Using this conceptual model, we can start to identify broad archetypes and trends in how businesses are approaching access to data and algorithms as a means of forming a competitive advantage.

Below, we lay out what this approach might look like and conceptualise the broad strategies that are characterised by areas on the matrix. While we have used our research and interviews to help infer these potential strategies, we have deliberately avoided classifying existing products and services. We hope, however, that this framing helps people understand some of the strategies being employed by businesses to create a competitive advantage through access to data and algorithms.

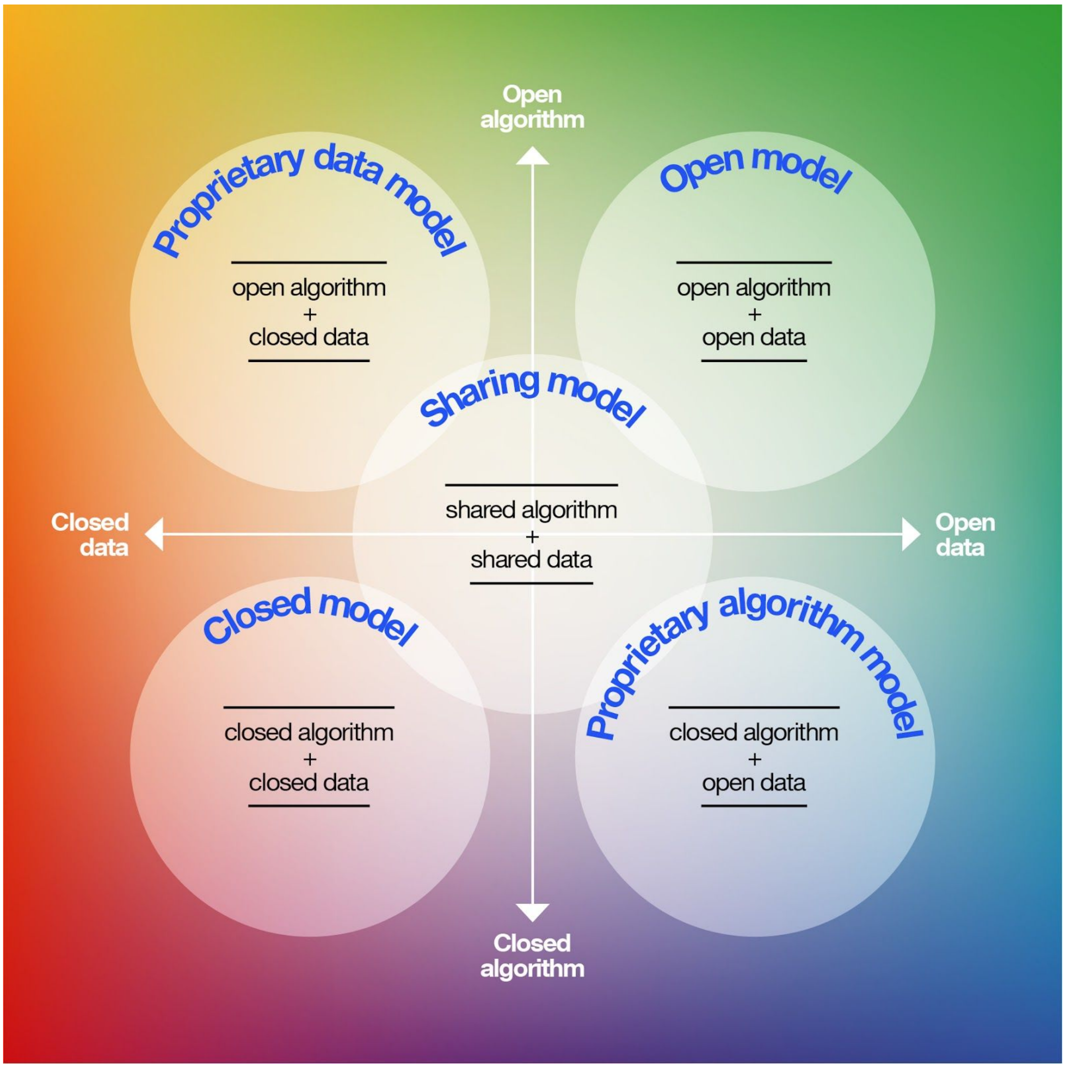
Open algorithm
Proprietary data model
open algorithm + closed data
Open model
open algorithm + open data
Sharing model
shared algorithm + shared data
Closed data
Open data
Closed model
closed algorithm + closed data
Proprietary algorithm model
closed algorithm + open data
Closed algorithm

Figure 2. The AI Business Model Matrix.

[19] Allison Linn (2016), 'Microsoft releases CNTK, its open source deep learning toolkit, on GitHub', https://blogs.microsoft.com/ai/microsoft-releases-cntk-its-open-source-deep-learning-toolkit-on-github

[20] TensorFlow (2018), https://www.tensorflow.org

Broad archetypes of AI business models

Proprietary data model: open algorithm + closed data

The archetypal business model near the top left of the matrix can be described as the proprietary data model. This approach is characterised by restricting access to data while providing wider access to AI algorithms.

Businesses employing this type of approach may aim to gain a commercial advantage by restricting access to data used to develop AI systems to prevent others from developing competing systems. At the same time, by using open source AI toolkits – or enabling access to their own algorithms and toolkits – companies are able to benefit from the latest advances in AI modelling, receive feedback from others in the AI community and encourage the development of complimentary services through open innovation.

Closed model: closed algorithm + closed data

The archetype positioned near the bottom left of the matrix can be described as the closed model. Businesses applying this approach restrict access to both the data and the algorithms under their control.

The benefit of this type of approach is that companies leverage their internal resources and develop their own skills and products, where the access to both data and algorithms constitutes their intellectual property. The downside is that it can be comparatively more difficult to have meaningful interactions with the wider AI community. Businesses that adopt a closed model tend to not only make fewer, less impactful contributions to the broader AI community, but also reap fewer of the benefits that engagement with the community can provide.

Proprietary algorithm model: closed algorithm + open data

The archetypal business model situated near the bottom right of the matrix can be described as the proprietary algorithm model – or what Maria Axente, a Digital Strategy Consultant at PwC UK, has called the 'intellectual property (IP) model'.[21] This type of approach is defined by greater access to data coupled with comparatively limited access to algorithms.

Companies that adopt this model tend to want to use their skills and expertise to develop proprietary algorithms that are more effective than their competitors and then allow outside companies with large amounts of untapped data to use their model, and access the outputs of that model, at a cost.

The advantage of this approach, especially for many startups and smaller businesses, is that it does not require them to control their own large data set – success is not predicated on the aggregation of vast amounts of data. In addition, there are benefits for the wider AI community, since greater sharing and opening of data encourages innovation.

Open model: open algorithm + open data

[21] This interview was conducted by Future Advocacy on behalf of the ODI.

The archetype close to the top right corner of the matrix can be described as the open model, where data and algorithms are both openly available.

Companies that employ this type of approach tend to want both elements openly accessible by the general public. This approach offers the most benefits in terms of innovation and speed of AI improvements, contributing to making society and the AI community better. However, choosing this strategy might require businesses to generate a competitive advantage through other means beyond restricting access to data or algorithms, such as skills or hardware.

Shared model: shared algorithm + shared data

In the centre of the matrix exist a range of shared models where access to data and algorithms are shared with select parties. Businesses employing these types of strategies tend to aim to derive competitive advantage from other facets of their business model, such as their skills and expertise. This might take the form of consultancies building bespoke AI systems for clients or licensing particular AI systems and associated services.

Outputs and other areas for competitive advantage

As previously stated, there are many areas in which products and services built on AI systems can compete. Beyond those that are common to all services, such as price or functionality, one remaining common facet for those built on AI systems is access to the outputs of AI systems. Access to these outputs can also be described on a spectrum, and this might impact how businesses attempt to create competitive advantage. For example, an organisation might provide access to both data and algorithms but limit access to the outputs, allowing them to sell the insights or functionality as a service. This may enable them to compete on the basis of access to hardware or skills, for example by running their system more responsively or by tailoring specific insights for clients.

On the whole there are many ways in which businesses attempt to create advantage over their competition by restricting access to some of their resources. However, there are many ways in which sharing and opening access to facets can provide tangible business benefits. This particularly applies in the case of digital assets such as algorithms and data. Businesses need to identify the approach that best works for them, depending on the context in which they are building an AI system.

The current state of play

At present, algorithms are shared or made open much more frequently than data due to concerns around personal data and commercial sensitivity.

The trend towards open algorithms and closed data

In general, companies see access to data as the key to gaining a competitive advantage when building AI systems and are choosing to restrict access in order to create a competitive advantage.

Many larger companies have started to set their sights on what Daniel Faggella of TechEmergence referred to as a 'self-feeding data ecosystem'.[22] Businesses with large siloes of high quality data are able to train the best AI systems; the best systems find the widest usage; the wide usage of these systems contributes yet more data to the siloes of the largest companies with the best AI systems. That data can then be used to improve existing models and create entirely new ones, creating a reinforcing cycle.

In addition, by collecting vast quantities of data as rapidly and as extensively as possible, organisations come to control what Gary Swart of Polaris Partners has called a 'proprietary data plume'.[23] Once enough data is collected – and the processes for data collection in real-time are efficient and well established – new companies attempting to move into the field have little option but to rely on the existing 'data plume' controlled by the early movers in the sector. This has been termed the 'winner-takes-all' model: access to an extensive enough dataset means that new efforts at collecting data – essentially, reinventing the wheel – are doomed to be uneconomical and therefore commercially disadvantageous. Indeed, according to Faggella, it is access to – or credible means for developing – a proprietary data plume that investors in AI companies value above all else when making investment decisions.[24]

This attitude to data is in contrast with how some algorithms and AI technologies are viewed by many businesses currently developing AI systems. Many AI technologies such as Keras, Apache Spark and Tensorflow are provided freely and openly, allowing them to be easily reused by other businesses.

The 'Open Algorithm + Closed Data' setup has led to great disparities in access to data across the AI sector. According to John Enevoldsen of Ocean Protocol – a company building a decentralised data exchange to help unlock data for AI systems – there is "a huge gap between the data haves and the data have nots" in the AI sector. It is the "big corporations and companies that have a lot of data", he explained, and not many of them are sure how to use it to build effective AI systems. On the other hand, AI startups "lack the relevant and high-quality data sets to actually train their models."

Dougal Featherstone from Frosha noted from experience that even in instances where people have expressed a desire to share their data with startups and SMEs, it is often not possible to access that data due to its being "ring fenced" inside major technology companies. For Featherstone, the key question to ask going forward is therefore, "How do you get that data out to companies like ours who would love to have it?"

[22] This interview was conducted by Future Advocacy on behalf of the ODI.
[23] Daniel Fagella (2017), 'Gary Swart on Defensibility and Scale for AI Companies', https://www.techemergence.com/garyswartdefensibilityandscaleaicompanies/
[24] This interview was conducted by Future Advocacy on behalf of the ODI.

Why are businesses taking this approach?

Businesses are increasingly choosing to provide access to their AI tools and algorithms, providing them under open source licences so that anyone can reuse them. It is increasingly the norm for businesses to publish code openly as a means of getting others to use it and help identify potential or existing issues, alongside other potential benefits.[25] In addition, developing AI systems in the open can have significant benefits for the development of AI systems in general, speeding up the adoption of new methods and ways of operating. This desire to drive progress in AI systems has been cited by many businesses as a reason for why they have made their AI toolkits open source.[26] A final reason for increasing access to algorithms is emerging around the idea of 'algorithmic accountability' – being able to understand how certain algorithms come to decisions.[27] Whether or not opening up models to interrogation can actually help hold algorithms to account remains in question, however – especially for certain types of neural networks. For this reason, it is not currently a big driver behind opening up algorithms but looks set to play an increasing role.

When asked why businesses are hesitant to share or open data, our interviewees tended to give two responses. First, they said the data in question is often sensitive, personal data that can be difficult to share for privacy reasons. Second, that it is often proprietary data that forms the basis of a company's competitive edge.

Roemer Claasen from Frosha expressed the first point quite succinctly. When asked whether having access to data that is currently siloed might encourage innovation in their sector, Claasen responded: "Of course we would be able to develop more business models out of that, of course, but as a citizen I would probably object to that. It's with good right there are some silos". This is a common approach to data about people within businesses: access to some data should be restricted unless the person it is about has given explicit consent. Besides the privacy issue, there are strong business reasons for controlling access to sensitive data about people – namely retaining the trust of customers and avoiding potential regulatory issues. This second reason makes many businesses risk-averse when it comes to sharing personal data. While it can be an admirable and justifiable approach, it can also be used by companies as an excuse to justify restricting access to what they perceive as proprietary data.

With regards to proprietary data, David Zomerdijk from Frosha followed his colleague's point about the valid reasons for siloing some personal data by explaining that while the sharing or opening of non-personal, commercial data would "cause innovation – a lot of it", companies resist because they "get some market advantage" of that data.

There are valid reasons for closing or limiting access to commercial data – for example when the data reveals something about the inner workings of a business, such as a profit and loss register. However, often businesses default to restricting access to data, even if there is potential value in sharing or opening that data. Businesses may keep some data closed, even if they are not using the data themselves – potentially losing out on capitalising on the value that it might have to others.

[25] Ibrahim Haddad (2018), 'Using Open Source Code', https://www.linuxfoundation.org/using-open-source-code
[26] Allison Linn (2016), 'Microsoft releases CNTK, its open source deep learning toolkit, on GitHub', https://blogs.microsoft.com/ai/microsoft-releases-cntk-its-open-source-deep-learning-toolkit-on-github
[27] Megan Rose Dickey (2017), 'Algorithmic accountability', https://techcrunch.com/2017/04/30/algorithmic-accountability
World Wide Web (2017), 'A Smart Web for a More Equal Future'
http://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf
OpenAI (2018), https://openai.com

According to our interviewees, businesses developing AI systems often do not see the value of opening the data they collect. Daniel Vila Suero from RECOGNAI, a startup building knowledge-powered predictive models, explained that in his experience, many – though not all – AI companies remain unconvinced that they will benefit from sharing or opening their data to others. He added, however, that while they resist opening or sharing their own data, many of those same companies do see advantages in the open data model in general, as they likely already benefit from other businesses and organisations opening up their own data.

Similarly, according to a number of our interviewees, some AI businesses do not see the value in using open data to train their models. Anne-Valérie Bach from Serena Capital – a top tier investment partner – explained that many companies feel that the types of data published openly by public sector organisations is often of poor quality, and would need to be extensively cleaned and combined with other proprietary data sets in order to prove useful.

> “ *Most of the time it’s a combination. Large companies or large organisations, they sit on a lot of data that they don’t even know what to do with, so usually it’s those people who need help in figuring out what to do but on the other hand, the start-up, if they only use open data they might not have the relevance and the quality they really need to train the algorithm, so usually it’s a combination.*
>
> ***– Anne-Valérie Bach, Serena Capital***

While it is true that some businesses developing AI systems do not see the value of opening their data or using open data, many AI systems are actually trained primarily or exclusively on open or publicly accessible data. For instance, it is possible to train natural language processing systems on text from open sources like Wikipedia[28] or publicly accessible sources like Twitter. A few of our interviewees developed their AI systems using primarily open or publicly available data. Frosha, for instance, combine data from several open sources to train the algorithms and are looking at using Twitter as a source of training data. Similarly, AI Poli – a company that combines natural language processing and machine learning to help business and governments assess the validity and relevance of news publications – regularly train their models on a combination of data made available by partners and publicly accessible data “scraped from the internet”.

It may be that people in the AI community do not perceive these publicly accessible corpora as ‘open data’, since some of them are not openly licensed,[29] but the wide usage of such data within the AI community demonstrates the value in making that data accessible.

Anne-Valérie Bach explained that since the business model of many companies developing AI systems is built around gaining exclusive access to better data than their competitors, open data sets – which by definition can be accessed by anyone, including competitors – do not give AI businesses a competitive advantage.
Both factors – the perceived lack of quality and the perceived lack of competitive advantage – likely contribute to the trend towards restricting access to data used to develop AI systems.

[28] Lingatools, ‘Wikipedia Comparable Corpora’, http://linguatools.org/tools/corpora/wikipedia-comparable-corpora

[29] Anna Scott (2017), ‘What is ‘open data’ and why should we care?’, https://theodi.org/article/what-is-open-data-and-why-should-we-care

Challenges related to the widespread siloing of data

Restricting access to data for building AI systems is stifling innovation and creating more biased AI systems

While many of the people we interviewed were sympathetic to arguments against sharing or opening data, they recognised some of the issues arising from this approach – not only for businesses, but for the economy and society as well. In particular, they focused on two major challenges related to the current treatment of data and algorithms within AI business models.

Stifling innovation and creating oligopoly

First, restricting access to data to retain a commercial advantage brings with it the risk that the development of AI systems will be dominated by a handful of large companies. Since access to large amounts of quality data is of paramount importance for developing AI systems, the monopolisation of data by a small cadre of businesses will mean that fewer startups, SMEs, and other businesses can turn their innovative ideas into new AI systems.

Michaele Veale commented that while it is unclear how the sector will develop over the next few years, there is a "natural tendency to monopolisation with machine learning systems." This is due in part to the realities of the 'self-feeding data ecosystem', identified in the previous section. Those with the most data are often able to train the best AI systems and those with the best AI systems are often able to collect the most data.

Veale cautioned, however, that such a setup was likely not sustainable. Speaking specifically about the role of deep learning in APIs, he explained: "Because deep learning just needs so much data it's very difficult to make a competitive market without legislation that would allow that data to be open".

The lack of a competitive market would hinder innovation and stall what is currently a thriving AI sector. It would stifle the energy and creativity that startups and SMEs contribute and lead, in effect, to a data oligopoly where access to data is concentrated in the hands of a few large businesses. This would not only hurt startups and SMEs, but large companies as well, since it would lead to a reduction in the diversity of problems to which AI systems are applied. Ultimately, it would lead to a reduction in the size of the AI market. Businesses should look to address these challenges head on rather than exclusively relying on government to tackle them through regulation.

If businesses fail to overcome these challenge, it will severely limit the ability of all businesses to build new innovative products and services based on AI systems. This includes those businesses that have built data oligopolies in certain areas, as they will be unlikely to gain access to certain types of data collected by their direct competitors.

Encoding unintended bias into AI systems

According to our interviewees, the second major problem with the current restrictions on access to data is that they will likely to lead to AI systems – in particular those based on machine learning – that contain more unintended bias in their models.

Although AI systems can often feel like black boxes that upon being fed with data (photos, preferences, interactions, etc.) magically return decisions, recommendations or pattern recognition, they are not magic – they are the creations of scientists, engineers, and programmers, and are subsequently trained to maturity on large amounts of data. The decisions and recommendations provided by an AI system are therefore a product of data and decisions made during the creation and training of that AI system.

Considering the crucial role data plays in training and operating AI systems, the datasets used to train those systems must be of sufficient quality and variety. An AI system trained on data gathered from an incomplete or unrepresentative section of the public, for instance, might develop racial, gender, or economic biases. For example, an image classification algorithm might do poorly at classifying images of minority ethnic groups compared with majority ethnic groups.[30] Along these same lines, biases present in the sampling or collection of data will likely become biases in the actual operation of any AI system trained on that data.

> “ *If you talk to machine learning people they will tell you, if you don’t have a rich data set you actually start discriminating people because there is bias in the data set.*
>
> ***– Sandra Wachter***

This type of bias may not have a huge impact on the efficacy of some systems – for instance, unintended bias in an AI system in charge of a home thermostat might only result in minor nuisances. But it might have a huge impact on people in other cases – for example bias in an AI system that screens job candidate CVs might affect the career prospects of underrepresented minorities. The impact of unintended bias may also vary by context, and might not be predictable by those building the system. In the thermostat example, imagine if the thermostat was used for regulating the temperature of a hospital room in winter, where the difference of a couple degrees could mean life or death. Bias in AI systems is not only potentially bad for people and society, but for businesses too, since a poorly functioning product can lead to a decline in consumer confidence.

Access to data from many different sources can help to achieve a more representative sample and can also help safeguard against AI systems reflecting any existing or historical biases related to the collection of data. In contrast, the current practice of limiting access to data could increase the likelihood that AI models will be trained on unrepresentative data.

Sandra Wachter, a research fellow at the Oxford Internet Institute (focusing on the legal and ethical implications of big data and AI), expressed this concern in her interview.

> “ *If you don’t know where the data comes from, you don’t know if the data is accurate, you don’t know if it’s actually complete. So having enough access to broad datasets could actually help to get rid of those biases.*
>
> ***– Sandra Wachter***

[30] Tom Simonite (2018), ‘When It Comes to Gorillas, Google Photos Remains Blind’, https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind.

She closed by expressing her belief that despite the importance of addressing this issue, it is something that “hasn’t really been communicated yet.”

From a business perspective, Daniel Vila Suero from RECOGNAI explained that in his estimation combining multiple datasets presents businesses with an opportunity to learn from more structured data. “It wouldn’t be a threat for AI companies,” he stated, “I think it’s more of a benefit.”

Part of the reasoning here is that combining data sets helps to make your data more representative and therefore better. Better data means better AI systems; better AI systems means better services; and better services means more customers. On the other hand, as Gil Arditi, head of machine learning at the ridesharing company Lyft, put it: “Bad data can lead to suboptimal decisions that actually kill competitive advantage.”[31]

On a related note, polling data suggests that a not insignificant proportion of people are concerned about the role such systems are likely to play in people’s lives in the future.[32] In such an environment, businesses developing AI systems will need to work especially hard with users to build trust in their AI systems.

While some advocate opening and sharing algorithms as a way of examining algorithms for inbuilt biases, many of those same people can overlook the opportunity to guard against encoded biases by opening and sharing data. At present, there is still a debate around whether it is possible to meaningfully understand certain AI systems – namely those dependent on deep learning – or the recommendations they offer, merely by interrogating the AI algorithm.

However, even in instances where it is not possible to fully interrogate an AI system, disclosing source-code and algorithms is an important step. The recent report on public scrutiny of automated decision-making by the Omidyar Network[33] noted that while opening and sharing algorithms may not always be enough to serve the needs of the public,

> “ *… transparency at different layers of a system can help external scrutinizers understand a system well enough to raise concrete concerns, guide further investigation, and propose potential remedies*

At the ODI we believe that a complimentary step in the right direction would be to open up and share data in the same way algorithms currently are. This would provide greater transparency at a different ‘layer of the system’ and would make it possible to examine the training data that may be equally – if not more – responsible for much of the inbuilt bias in AI systems. However, this approach is also likely to be contested in terms of its

[31] John Koetsier (2017), ‘Artificial Intelligence and Commerce: 4 Steps To Bringing AI To Market’, https://www.forbes.com/sites/johnkoetsier/2017/10/24/4-steps-to-bringing-ai-powered-products-to-market-how-lyft-builds-intelligence-into-product/#698155421b66

[32] The Royal Society (2017), ‘Public Views of Machine Learning’, https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf

[33] Omidyard Network (2018), ‘Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods’, http://omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods

ability to fully reveal unintended encoded bias, especially as it is currently relatively untested compared to other methods.

Sharing and opening data could not only help safeguard against the dangers of unintentional encoded biases, but encourage innovation and benefit businesses. Failing to do so would be bad for society, bad for people, and bad for business.

The role of regulation, competition and trust

As people gain more rights to exercise control over data about them, businesses need to think about the potential opportunities and impact these rights might have on the design of AI systems

When asked what the future might hold, the people we spoke to identified a number of trends they believe have the potential to significantly impact current AI business models and practices around data. These trends have the potential to both positively and negatively impact the development of future AI systems.

Trust and a shift towards greater personal control of data

One trend identified by our interviewees was a general move towards greater emphasis on transparency and trust. Matthew Sheret of Projects by IF referred to his experience of working with clients who were not predominantly concerned with the acquisition of new data, but rather with developing AI systems that are more open and transparent.

> “ *What is important is trust – this is what a long term relationship between services and users is built on.*
>
> ***– Matthew Sheret***

His comment suggests there is an increasing commercial imperative in being trustworthy in interactions with potential and existing clients. Sheret defines trustworthiness as “building services in a way that makes it clear what’s going on to users and that can be unpicked by regulators and independent groups”.

Building trust will likely become increasingly important for businesses over the coming years as people are given greater control over data about them. Indeed, quite a few of our interviewees mentioned the imminent arrival of the General Data Protection Regulation (GDPR) in May,[34] and what they saw as a general shift towards a data economy where people are given greater control over the sharing and usage of data about them.

[34] EUGDPR (2018), https://www.eugdpr.org

General Data Protection Regulation (GDPR)[35]

GDPR impacts how companies collect, use and share data about their users by giving people rights over data about them. It applies to all businesses processing data about citizens of the European Union, even if the business is headquartered outside of the EU. Though GDPR creates a few new rights around data, it is only the latest iteration of data protection regulation affecting citizens of the EU.[36] Under GDPR, people have specific rights over data about themselves, including over how or whether it is collected, managed, shared, and deleted. Of particular interest to businesses building AI systems is the set of rights given to people regarding automated decision-making.[37]

It is unlikely that everyone will exercise these new rights immediately – recent research in the retail sector, carried out by the ODI, found that only 25% of shoppers were likely to exercise their right to portability.[38] However, there is potential that more people might choose to exercise these rights in response to a loss of trust in AI systems or the businesses using them. More importantly, new products and services, or features for existing services, might make use of these rights and these may enable people to exercise these rights easily. For example, an application might ask for permission to access data about your spending habits to recommend some other service, and if granted, exercises your right to access data for you.

If businesses and organisations lose the trust of their users, they may find that their users are more reluctant to share their data or allow companies to use their data to train new models. Less access to data could result in less functional AI systems – this would not only hurt the individual business, but could hurt the wider AI business community by leading to a drastic reduction in relevant, accessible training data.

An opportunity rather than a threat

Seen from a different perspective, businesses that help their customers balance their desire for improved services with the desire for privacy, will preserve trust. Thus, businesses and organisations developing AI systems need not see this trend toward greater personal control of data as a bad thing. Indeed, it could be a big opportunity for innovation.

Many of our interviewees argued that businesses should see this trend toward greater personal control of data as a business opportunity to be approached proactively and positively. Michael Veale, for instance, felt that GDPR should serve as a wake up call to businesses. Rather than seeking to avoid or fight against these realities, Veale advised businesses to say, "Okay, we need to think about the business model for this – we can't pretend we're all closed all the time".

If a business designs a service that builds trust between them and their users by simplifying the process of exercising rights over data and making it easier for users to understand how their data is being used, they may attract more users. That business will have created a market and gained a commercial advantage over businesses that

[35] Information Commissioner's Office, 'Guide to the General Data Protection Regulation (GDPR)', https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr.

[36] ICO (September 2017), 'Overview of the General Data Protection Regulation', https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr.

[37] Information Commissioner's Office. Rights Related to Automated Decision Making Including Profiling', https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling

[38] The Open Data Institute (2017), 'The EU GDPR Opportunities for Grocery Retail', https://drive.google.com/file/d/1A_5gL2504AaXWEvig7HZSWIex13oMIzs/view.

are slow to respond to the new realities of a future where people are given more control over data about them.

The importance of this trend in driving future business practice was also highlighted by Maria Axente. While many AI businesses currently see the collection and aggregation of data as the key to gaining a competitive edge, Axente postulated that the prospect of shifting control of data towards consumers and away from companies will increase the attractiveness of the 'closed algorithm + open data' model.[39] If companies find that they are no longer able to restrict access to data they have collected about people, then they may decide that the best way to add value is to sell access to algorithms that allow analysis of customers' datasets – i.e. the 'proprietary algorithm' model.

Sandra Wachter felt that a shift towards giving people greater control over data about them would be a boon for businesses developing AI systems. "I actually think it will enable innovation", she stated during our interview: "it redirects markets and refocuses research on something new". She was confident that by applying themselves to the new realities around privacy, transparency and accountability, AI researchers would be able to come up with "tools that will address the market". Michael Veale felt similarly and added that this shift was "quite an important move to break down some of the monopolies" that might otherwise form.

Wachter warned that refraining from collecting any data that might be deemed personal or sensitive out of fear of running afoul of data protection regulations, would be the wrong approach.

> “ *It is counter-intuitive because you might want to say, 'because of data privacy considerations you want to collect as little data as possible', but if you do that you actually end up with more discrimination than you would otherwise.*
>
> ***– Sandra Wachter***

Her reasoning was that by restricting the collection of data, businesses will almost guarantee that AI systems will have ingrained biases because those systems will be trained on limited data sets. She closed by noting: "I think there is some tension here that needs to be resolved in the future." To build effective, unbiased (and therefore more profitable) AI systems, businesses will need to have access to as much data as is legally and ethically possible.

Government intervention and regulation

Over the coming years, businesses developing AI systems will need to decide how to respond to the trend towards people having greater control over data about them. By responding proactively and positively, businesses developing AI systems can turn a potential threat to their business model into an opportunity for innovation – both on an individual basis and as a sector.

On a practical level, businesses that adopt good practices around GDPR or transparent algorithmic decision-making will lessen the likelihood of legal challenges and potentially hefty fines.[40]

[39] This interview was conducted by Future Advocacy on behalf of the ODI.

[40] Matthew Sheret (2017), 'Designing for Trust', https://projectsbyif.com/blog/designing-for-trust

On another level, by moving to grant wider access to data they use, businesses developing AI systems may be able to avoid potential anti-monopolistic government intervention. If, over the coming years, governments and regulatory agencies find that a data oligopoly is hindering innovation in the AI community, they may move to challenge those monopolies, especially if governments themselves find that the efficacy of their own services is being hindered by a lack of access to data caused by companies choosing to restrict access to data.

Examples of similar moves in the past include ‘United States v. Microsoft Corporation 253 F.3d 34 (D.C. Cir. 2001)’, a U.S. antitrust law case wherein Microsoft’s market dominance was judged to constitute a monopoly, as well as the European Commission’s recent €2.42 billion fine of Google, alleging that the company “abused its market dominance as a search engine by giving an illegal advantage to another Google product, its comparison shopping service.”[41]

The landscape for businesses developing AI systems is changing. By responding positively and proactively, businesses can capitalise on these changing circumstances.

[41] Joel Brinkley (2000), ‘U.S. VS. Microsoft: the Overview; U.S. Judge says Microsoft Violated Antitrust Laws with Predatory Behavior’, http://www.nytimes.com/2000/04/04/business/us-vs-microsoft-overview-us-judge-says-microsoft-violated-antitrust-laws-with.html
European Commission (2017), ‘Antitrust: Commission fines Google €2.42 billion for abusing dominance as search engine by giving illegal advantage to own comparison shopping service’, http://europa.eu/rapid/press-release_IP-17-1784_en.htm.

Conclusion: opening and sharing data

To overcome the challenges arising from restricting access to data and capitalise on the opportunities as people exercise more control over data about them, more data must be shared with more businesses, in a way that retains the trust of people while promoting innovation

To grow the AI sector in a way that benefits businesses, governments, and people, we will need to ensure that the system respects commercial interests while actively giving users control over data about them and enabling users to make decisions about how data about them is used and shared. Businesses should seek explicit permission from their customers and others the data might be about when opening or sharing data about them. Giving users granular control about how their data might be shared and explaining the risks around it will enable them to make informed decisions.

Sharing and opening data can help support business objectives as well as benefiting society – for example it can be used to build up a platform for engaging new and existing customers, enable the production of complimentary services, and encourage open innovation around strategic challenges.[42]

Businesses need to consider how they will provide greater access to data in a way that people will trust. In all cases, this will rely on clear communication of not only the methods being used but also the benefits this might have.

Publishing open data

Not all data that is useful for building AI systems is personal data or commercially sensitive. In many cases, there are few limitations on providing wider access to this data. To maximise the potential value – to the business, the AI sector as a whole, and society – businesses should publish this data under an open licence as open data. There is clearly a desire for more publicly available datasets within the AI sector – 72% of respondents to a recent Digital Catapult survey of machine learning companies wanted the catapult to provide access to public datasets.[43]

> “ *Unsurprisingly, data was the most in-demand item, specifically access to new public (or private, sandboxed) datasets for training machine learning models.*

[42] The Open Data (2016), ‘Open enterprise: how three big businesses create value with open innovation’, https://theodi.org/article/open-enterprise-how-three-big-businesses-create-value-with-open-innovation
[43] Digital Catapult (2018), ‘Machines for Machine Intelligence’, https://www.digitalcatapultcentre.org.uk/machines-machine-intelligence-report

As highlighted earlier, the majority of existing open data sources come from public sector organisations, and there have been questions about the quality of that data. However, increasingly, the private sector is making data available openly and this trend looks set to continue. As Michael Veale argued, “the tide is going towards data being more open” and businesses will therefore have to become “more open with their data” going forward.

Publishing open data can be relatively easy for businesses. All they need to do is make a dataset or expose an API endpoint in a way that is accessible on the web and make it clear that it is published under an open licence. The licence is key, enabling users of that data to easily understand what they can do with the data. The licence should meet the requirements of Open Knowledge’s Open Definition.[44] Best practice recommends the use of unedited standard open licences, such as those provided by Creative Commons.[45] By applying these licences to existing public sources, data users can be more confident in using the data to develop innovative products and services.

Businesses should also include detailed metadata about the data as part of best practice, enabling users to understand what the data is, its source and potential limitations. The ODI’s Open Data Certificates[46] and guides provide significant guidance on open data publishing, as well as helping to make data discoverable and usable.[47] Businesses could also choose to use existing open source or third party tools to publish data, including the ODI’s Octopub tool[48] or a platform like data.world,[49] which make the process easier. For high-volume data, businesses looking to recover some of the costs of publishing can consider freemium models for data publishing such as those offered by OpenCorporates.[50]

For data that does contain content about people, there are potential strategies that businesses can use to publish this data. Anonymisation and aggregation techniques can be used to lower the risk of individuals being identified.[51] Explicit consent to share details about people can also be sought for certain types of data, for example people donating money to particular organisations might give permission for their contribution to be published openly. Creating a system where explicit individual permissions can be obtained and combined with clear licensing terms for the use of that data by other organisations could help companies developing AI systems avoid breaching the trust of people when using publicly available data sets to train their AI systems. In some rare cases, it might be possible to publish data about people without consent if its already widely available, for example data about public figures or people that have been deceased for a long period of time.

Of course, not all data can be opened – especially where people choose not to share data about them. However there are emerging techniques for sharing granular data without revealing details about individuals or commercially sensitive data, especially around machine learning. Synthetic data is one of many techniques for de-identification

[44] Open Definition, Open Definition 2.1, http://opendefinition.org/od/2.1/en

[45] Creative Commons, ‘About the licenses’, https://creativecommons.org/licenses

[46] The Open Data Institute, ‘The mark of quality and trust for open data’, https://certificates.theodi.org/en

[47] The Open Data Institute, ‘Data publishing and use’, https://theodi.org/topic/data-publishing-and-use

[48] Octopub: https://octopub.io

[49] Data World: https://data.world

[50] For more information about OpenCorporates, see: https://opencorporates.com/api_accounts/new.
For more information on potential business models around open data see: The ODI (2017), ‘Data entrepreneurship: exploring successful business models with open data’, https://theodi.org/article/data-entrepreneurship-exploring-successful-business-models-with-open-data

[51] We write ‘lower the risk’ since arguably it may not always be possible to fully anonymise data. For more information, see: https://ico.org.uk/for-organisations/guide-to-data-protection/anonymisation

of data.[52] Synthetic data is created artificially from real data to maintain the overall statistical understanding while changing the individual records. The ability to train AI systems with this kind of data would substantially alleviate concerns about sharing potentially identifiable or commercially sensitive data. There have been some documented success in academia[53][54] and the commercial sector is seeing the emergence of startups trying to commoditise the practice.[55]

Sharing data

Our interviewees commented frequently on the potential for businesses to share data within mutually beneficial partnerships and collaborations. The key, according to Daniel Staff of Digital Catapult, would be to match businesses without access to data with data rich businesses in ways that ensure both parties receive something of value from the partnership. On one side, Staff explained, “you have a company that’s sitting on a lot of data but not necessarily leveraging it because they’re not aware of new techniques and new research”, while on the other you have “a small startup that has the knowledge but not necessarily the data”. In such a collaboration, the company with a large data set would receive the technical expertise needed to unlock valuable insights from their data, contributing to that company’s competitive advantage. Conversely, the startup would receive access to large datasets that are crucial for training and developing innovative AI systems. Digital Catapult is one organisation helping to create this type of collaboration through open innovation activities and other programmes such as the Machine Intelligence Garage.[56]

There is great potential for partnerships and collaborations that involve not only private sector companies, but government bodies, universities, non-profits, and public sector organisations as well.

> “ *We need to explore how to encourage better data equity so that we have some stronger sharing protocols between companies like Facebook and all the social science labs in the universities around the world. A lot of this is very obscure to us because we don’t have access because their data is extraordinarily valuable but there ought to be some way of managing large scale data sharing efforts that are in everybody’s interest.*
>
> ***– Sir Nigel Shadbolt***

There are a number of difficulties in realising these types of mutually beneficial collaborations, not least of which is identifying suitable partners. One approach being

[52] Personal Data Protection Commission (2018), ‘Guide to Anonymisation Techniques’, https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf

[53] Adam Kortylewski et al (2018), ‘Training Deep Face Recognition Systems with Synthetic Data’, http://search.arxiv.org:8081/paper.jsp?r=1802.05891

[54] Stefanie Koperniak (2017), ‘Artificial data give the same results as real data — without compromising privacy’, http://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303

[55] Neuromation: https://neuromation.io/en

[56] Machine Intelligence Garage: https://www.migarage.ai

taken for finding innovative partners is adopting open innovation approaches. In some cases this is defined as a challenge, where a company with a large data set and, in some cases, a well-defined problem puts out an open call for startups and SMEs to develop innovative solutions to that problem. This can be done by releasing the data up front – for example the Kaggle competition platform enables big businesses to host open data science competitions around specific data and specific problems, and anyone can enter.[57]

Alternatively, businesses can put out open calls for partners and then enter ongoing data sharing arrangements. For example, the EU-funded Data Pitch project pairs corporate and public-sector organisations that have data with startups and SMEs that work with data, as well as providing lots of support for these collaborations. Businesses could take either approach or a mix of the two depending on the outcomes they are looking for.

Whether partners are identified directly or through open innovation approaches, the next step for these businesses will be to build trust between the various parties, and finding adequate ethical, legal and technical means of sharing data between them. This is particularly challenging when personal data such as journey or health data is involved, but even non-personal commercial data can be difficult to share.

Dougal Featherstone of Frosha explained that in his experience it is often not possible for SMEs like Frosha to gain access to data even when people and companies are willing to share it with them. With an eye to the future he asked, “How do we build that bridge?”

Data trusts and exchanges

One potential way of bridging the gap between parties that want to share data but currently find it difficult to do so, is through data exchanges, marketplaces or trusts. Each of these types of arrangement aim to provide the technical and legal infrastructure to allow businesses and other organisations to share data in a secure manner. The aim is to provide mechanisms that facilitate sharing between a number of different parties in a more frictionless way than lots of individual data sharing agreements.

In our interview, Nick Allott of nquiringminds – a data analytics company that hosts a data sharing platform – noted that there is “massive potential value in being able to share raw data and analyse streams of data between parties.” Value, he clarified, for large companies, SMEs, startups, public sector organisations, and universities alike.

Similarly, John Enevoldsen explained that one of the reasons why Ocean Protocol was started was to help AI businesses collaborate more freely by bridging the gap between the data haves and data have nots

According to Allott, one of the main reasons data is currently not shared between businesses and organisations is that it is difficult to trust other organisations with personal or commercial data. The goal of the data exchange operated by nquiringminds is to “afford certain controls, checks and balances and assurances that lubricate that data flow”. “The key to AI is access to data,” he explained, “the key to unlocking that data is having a secure platform that allows your customers to have confidence in it and share data between different people”.

Each of these types of arrangement have the potential to facilitate wider data sharing – which one is most appropriate will depend on the situation. Data trusts which are governed by collaboration rather than through financial incentives may offer a means for businesses to create significant, ongoing collaborations.

[57] Kaggle (2018), https://www.kaggle.com/competitions

Recommendations to businesses

Businesses need to identify the costs and benefits of making more data available as part of the AI systems they are building. They need to understand the potential implications of failing to increase access to data around stifling innovation and encoding unintended bias. In light of the current trend towards people having greater control over data about them, businesses need to design AI systems that create trust. If they do not, companies risk losing access to data, either from individuals withdrawing consent, competitors siphoning users, or through government interventions in the market. This being the case, businesses should:

i) Examine their current approach to data as part of their business model

Businesses should aim to understand how they are currently creating a competitive advantage in terms of access to data and algorithms. They may choose to adapt their current approach to data, exploring how they can improve access to data and realise the associated benefits.

ii) Consider what data might be made open

Businesses should identify where, if at all, they can publish data openly. This is subject to obtaining explicit consent if that data is about people, and understanding what de-identification techniques could be used to avoid the identification of individuals.

iii) Explore the options around creating data trusts

Businesses should examine the potential for using data trusts to share data securely between partners. If this is personal data, this is subject to obtaining explicit, informed consent from the people who the data is about.

In addition, there is still much more research to be done on how businesses can share and open data to support the development of the AI sector. This might include work by government and other organisations on how to encourage, facilitate and possibly incentivise greater collaboration and sharing of data within the AI sector.

At the ODI we will begin to explore some of these issues over the next year, specifically emerging methods of de-identification of data about people and the potential for data trusts to encourage collaboration within the AI sector.

> *It isn't a perfect world in terms of data and there's plenty more still to do. And right at the heart of that are questions of: What data should be open? What should be shared? What should stay closed? What's personal and what's not? That whole data spectrum discussion is so important still. The slight worry is that people may think: 'Data has been solved – there's just loads of it about, so we can worrying about that'.*
>
> ***– Sir Nigel Shadbolt***

Methodology

For this research, we commissioned Future Advocacy to produce an overview of the role of artificial intelligence and machine learning in business models and policy frameworks.

They conducted interviews and consulted academic publications, public policy reports, and news publications in order to explore the role of AI, machine learning and data in emerging business models. We then conducted a landscape review which involved extensive research of the current use of AI in the commercial sector.

We identified a list of startups and large firms using AI in their business, which we analysed and categorised according to the way that AI was deployed within each business. A number of businesses were selected for targeted, in-depth interviews based on how they turned AI into a commercial advantage.

We conducted 17 interviews with participants from a range of backgrounds, striving for a balance between people from academia, research institutions, startups, SMEs and venture capital. Our interviewees were mainly from the UK (13), though we also interviewed one person each from France, the Netherlands, Spain and the United States of America.