Written evidence submitted by Google (ADM0016)

1. Introduction

1.1 Google would like to thank the Committee for the opportunity to provide a submission about this important subject.

1.2 Algorithms are a longstanding concept in mathematics and computer science. In the broadest sense, an algorithm is simply a set of instructions. Cooking recipes, instructions on how to play a game, and walking directions are all everyday illustrations of what could be called algorithms. In essence, algorithms are just a way to mathematically represent a structured process for carrying out a task, no different to the often complex rules and step-by-step workflows that permeate existing decision-making processes. For example, in the public sector, to calculate eligibility for benefits or assess immigration status; in the private sector, to decide whether to grant a mortgage or when to stock store shelves. Algorithmic systems have long been ubiquitous—from anti-lock braking systems improving safety, to ATMs that make banking more convenient.

1.3 However, most people associate algorithms more narrowly with computing—the guidelines followed by a machine to make sense out of data. The format of these guidelines can vary dramatically from a list of simple "if X, then do Y" rules as in traditional programming to the latest methods being developed in the field of machine learning. It is the latter, computing-focused definition that we have adopted in our submission.

1.4 While algorithms have been used in computing for decades, they have taken on a growing importance in various parts of the economy and society in the last decade or so in line with the spread of computers as a business support tool. Clearly, the kind of scrutiny applied to decision-making in general should also apply to computer-supported decision making. In this sense, the issues are not new. Even challenges associated with integrating computers into the decision making process are a longstanding field of research—for example the field of human-computer interaction[1] (HCI) or psychological research into "algorithm aversion"[2] (such as the notion that if an algorithm errs, people lose confidence in it far more quickly than if a human had erred).

[1] https://www.epsrc.ac.uk/research/ourportfolio/researchareas/hci/

[2] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2466040

1.5 This submission is focused on how algorithms are used in decision-making, with particular reference to public policy. It outlines:

- How we understand the benefits from algorithms, and their likely future role in the work of businesses, public bodies, and the government.
- What we believe to be good practice in terms of providing safeguards around the use of algorithms.
- Consideration of existing policy proposals around the governance of algorithms.

1.6 In this submission, we set out that:

- Algorithms, particularly with the development of artificial intelligence (AI) and machine learning (ML), are powerful tools that could provide crucial help in advancing science and research, improving access to medical care, and tackling some of our most pressing global challenges in the environment, transportation, and beyond, as well as driving smarter solutions to everyday problems. Increased use of algorithms in the public sector, implemented with care, could deliver real improvements for the public.

- It is important to maintain an environment in which innovation is not stifled and users are protected. For example, pushing for 'full transparency' to reveal the raw code of a search engine would help spammers and people trying to game the system as well as conflicting with long-standing legal protections for trade secrets. Instead, we believe it is more valuable for users to understand the inputs that go into algorithm design and how they help achieve desired outputs.

- At the same time, it is vital to promote public understanding about the purpose of algorithms and their responsible development and application. At Google, we strongly believe in sharing good practice in order to ensure that the benefits of AI are spread as widely and as fairly as possible. This is why we established the People and Artificial Intelligence Research initiative (PAIR), with its focus on how AI systems must be people-centric.

- Bias in algorithms can be a risk—but it should also be remembered that bias might already exist in many existing non algorithmic processes and well-designed algorithms could reduce bias and make processes fairer. Good practice at Google in identifying and eliminating potential bias within AI includes scrutinising initial data-sets and inputs in order to eliminate bias, continually testing and encouraging feedback, and working with civil society.

1.7 Above all, we believe that algorithms are fundamental to driving social and economic progress, but the use of algorithms should not be removed from political debate or democratic oversight. We welcome the fact that the Committee is holding this inquiry.

Responses to inquiry questions:

2. The extent of current and future use of algorithms in decision-making in Government and public bodies, businesses and others, and the corresponding risks and opportunities

2.1 Google has relied on algorithms[3] since its inception to power everything from Search to Translate, helping millions of users to access a world of information. Of the myriad opportunities algorithms afford, this submission focuses on two: inclusion and security at scale.

2.2 Increasingly advanced algorithms enable new products and features that can make information more accessible, and our society more inclusive. Since 2006, Google has worked to expand the number of videos captioned on YouTube, increase the quality of those captions, and add new features and languages. One key component of this work is the use of algorithms which can automatically add captions to videos by analysing speech audio and converting it into text. As of February 2017, we have captioned more than 1 billion videos[4], with automatic captioned videos receiving more than 15 million visits per day. These developments are helping the millions of members of deaf and hard of hearing communities and others around the world access content previously unavailable to them.

2.3 We are increasingly using AI and ML to power our algorithms and deliver benefits to users. In our products, ML advances have boosted efforts ranging from user protection—with improved spam and malware filters—to enhanced accessibility, through stronger voice recognition. For example, using a method[5] that learns language from patterns in bilingual text, Google Translate[6] translates more than 100 billion words a day in 103 languages. AI has been key in helping us achieve significant breakthroughs with speech recognition[7], reaching near-human levels of accuracy. With Google Photos[8], you can search for anything from “hugs” to “border collies” because the system uses our latest image recognition system to automatically categorise objects and concepts in images. We also recently announced that ML helped optimise system settings to cut energy consumed cooling our data centres by up to 40%.[9]

2.4 Our library of algorithms goes beyond powering and improving our core product offerings. Algorithms are also a critical building block for ensuring that users worldwide can browse safely. For example, Google Safe Browsing analyses billions of URLs per day to find unsafe websites. These alerts are surfaced on Google Search and Chrome, Firefox, and Safari browsers and webmasters, updating them on the thousands of unsafe websites discovered every day. The upshot is that the scale offered through algorithms can enable better protection to users from malware and phishing attacks online.

---

[3] https://www.google.com/search/howsearchworks/algorithms/
[4] https://youtube.googleblog.com/2017/02/one-billion-captioned-videos.html
[5] https://www.youtube.com/watch?v=_GdSC1Z1Kzs
[6] https://translate.googleblog.com/2016/04/ten-years-of-google-translate.html
[7] http://fortune.com/2017/05/18/google-speech-recognition/
[8] https://googleblog.blogspot.com/2016/03/smarter-photo-albums-without-work.html
[9] https://www.blog.google/topics/environment/deepmind-ai-reduces-energy-used-for/

2.5 The combination of increasingly advanced algorithms, data, and cheaper computing power means that the benefits of algorithmic decision-making will become ever more broadly distributed and that new use cases will continue to emerge. This encompasses apps and services that can improve our daily lives, but also projects that can achieve broader societal goals. For example, last year Google showed how ML can make the diagnosis of diabetic retinopathy[10]—one of the fastest growing causes of blindness worldwide—more broadly accessible.

2.6 Within the government and among public bodies, leveraging modern algorithmic breakthroughs in machine learning to enhance decision-making remains a nascent but growing phenomenon. At its best, algorithmic decision-making can drive efficiency and help lower costs, increase the impact of the provision of government services, and improve the detection of fraud. The UK Serious Fraud Office has used algorithms to augment the ability for investigators[11] to identify evidence of wrongdoing in a recent case. Surtrac[12], a collaboration led by the Carnegie Mellon University Robotics Institute, leverages algorithms to enable cities to better manage traffic and reduce emissions. We believe many more applications will be identified in the years to come and encourage further examination of how government can use algorithm-driven decisions to be more effective, responsive, and inclusive.

2.7 At the same time, we join others in emphasising that these technologies be responsibly designed and implemented—particularly in the public sector. Improvements in cost and efficiency unlocked through algorithmic decision-making should be balanced against the need to ensure equality, accountability, and democratic participation in the creation and provision of public services.

2.8 Moreover, we do not believe that algorithmic decision-making will always be the optimal means of resolving challenges in public administration—they aren't a panacea or a catch-all solution. These technologies do not remove the importance of public consent and democratic participation, and in some cases it may well be paramount to preserve the deliberative, context-sensitive decision-making that having human involvement can provide.

2.9 As with earlier waves of technological advancement, social values remain important. We believe that convening and supporting robust discussion and collaboration between technology experts and civil society is a key role that government can play as it seeks to fully benefit from these advances. It is important to consider how we can maximise the positives of emerging technologies for society whilst minimising potential harms.

---

[10] https://www.blog.google/topics/machine-learning/detecting-diabetic-eye-disease-machine-learning/

[11] https://www.ft.com/content/55f3daf4-ee1a-11e6-ba01-119a44939bb6

[12] http://www.surtrac.net/

2.10 With this in mind, earlier in 2017, we launched the People + Artificial Intelligence Research Initiative (PAIR)[13]. The goal of PAIR is to make "people and AI partnerships productive, enjoyable and fair." PAIR is devoted to advancing the research and design of people-centric AI systems, covering the full spectrum of human interaction with machine intelligence, including:

- Everyday users: How might we ensure AI and ML is inclusive, so everyone can benefit from breakthroughs in AI? Can design thinking open up entirely new AI applications? Can we democratise the technology behind AI?
- Engineers and researchers: AI is built by people. How might we make it easier for engineers to build and understand ML systems? What educational materials and practical tools do they need?
- Domain experts: How can AI aid and augment professionals in their work? How might we support doctors, technicians, designers, farmers, and musicians as they increasingly use AI?

2.11 Overall, PAIR aims to conduct fundamental research, invent new technology, and create frameworks for design in order to drive a humanistic approach to AI while being as open as possible: building open source tools that everyone can use, hosting public events, and supporting academics in advancing the state of the art. More about PAIR, including demonstrations of several of the tools they have open-sourced, can be found in this video.[14]

3. Whether 'good practice' in algorithmic decision-making can be identified and spread, including in terms of:

- **The scope for algorithmic decision-making to eliminate, introduce or amplify biases or discrimination, and how any such bias can be detected and overcome;**

3.1 Algorithms are tools that can be used to benefit consumers, society, and the economy but, like any tool or technology, their use can cause harms too, including creating or exacerbating societal inequalities. We are aware that potential harms exist in the use of algorithms and ML, which is why we are working hard to help eliminate them at an early stage. For example, as part of the PAIR initiative we're conducting research into how emerging ML systems can ensure fairness and equality of opportunity, and working on visualisations and interactive explanations to help people understand these critical issues such as this video[15] highlighting the challenges of human bias in ML.

3.2 As a recent ProPublica[16] investigation into ML used by the judicial system in the US illustrated, partial or biased data can produce discriminatory results as ML algorithms draw incorrect inferences from the examples they are trained on. Equally, by focusing on more

[13] https://ai.google/pair/
[14] https://youtu.be/QS-G6CwYGUY
[15] https://www.youtube.com/watch?v=59bMh59JQDo
[16] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

objective criteria, ML might help reduce or avoid discrimination. In his article, Equality of opportunity in supervised learning[17], Google researcher Moritz Hardt looked at the short term questions of bias and discrimination. In the article Hardt points to the need for improved tools for diagnosing these failures, as well as the need to avoid data gaps[18] where the dearth of good data[19] can make the use of ML problematic.

3.3 Google takes these issues very seriously and shares the goals of the committee to demonstrate and spread good practice throughout the development and use of algorithms used in decision making. There are four components to our approach:

- Research and develop concrete tools to maximise positive impacts
- Understand and tackle underlying factors to problematic outcomes
- Establish a framework to guard against bias and demonstrate accountability
- Explore how algorithms can be used to positively address bias and inequality

Research and Develop Concrete Tools

3.4 Rigorous research can provide a path towards answering the societal questions that increased use of sophisticated algorithms raise. Research should be interdisciplinary, but computer science provides a unique opportunity to develop concrete technical mechanisms and frameworks that can be deployed to diminish or eliminate bias. We've been conducting a great deal of work in this area, including our published research on equality of opportunity[20], adding fairness constraints into training[21], designing transparent machine learning[22], designing fair auctions[23], and various techniques to detect, debug, and analyse machine learning systems.

3.5 Ultimately these tools and mechanisms will only get us so far. Mathematical and technical constraints force tradeoffs that practitioners and policymakers will need to resolve between accuracy and interpretability or fairness. Continued investment in research and development of tools from the public and private sectors for immediate challenges should inform those decisions.

Understand and Tackle Underlying Factors

3.6 No system is infallible, and errors will occur. The myriad of issues at hand are complex in terms of both their causes and impacts, and thus will not be remedied by one single solution. Therefore, better understanding of the causal or contributing factors is critical to designing effective methods to reduce or eliminate the bad outcomes algorithms can produce.

---

[17] https://research.googleblog.com/2016/10/equality-of-opportunity-in-machine.html
[18] http://www2.datainnovation.org/2014-data-poverty.pdf
[19] https://medium.com/@melindagates/to-close-the-gender-gap-we-have-to-close-the-data-gap-e6a36a242657
[20] Equality of Opportunity in Supervised Learning: https://arxiv.org/abs/1610.02413
[21] Satisfying Real-world Goals with Dataset Constraints: https://arxiv.org/abs/1606.07558
[22] Monotonic Calibrated Interpolated Look-Up Tables: https://arxiv.org/abs/1505.06378
[23] Fair Resource Allocation in a Volatile Marketplace: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2789380

3.7 Take, for example, two studies examining algorithm-powered advertising networks. Researchers at Carnegie Mellon published a study[24] with experiment-based observations, arguing that Google's advertising networks showed women fewer instances of ads for high paying jobs than males, although further investigation[25] from researchers at MIT and the London Business School found that the most likely cause was the economics of advertising and how advertising campaigns are structured. Specifically, the problem was not bias against women, but rather that women were less likely to receive instances of ads for high paying jobs because they're a more expensive demographic to advertise to and an ad optimised for cost-effectiveness would yield less impressions.

3.8 Data gaps[26] are another factor. Just as many have expressed concern with the digital divide, we need to ensure there is not a data divide. As recent cases make clear[27], a lack of good data, or poor quality, incomplete, or biased data sets are problematic in the practices and biases they perpetuate and can potentially produce inequitable results in algorithmic systems. "Growing the artificial intelligence industry in the UK"[28], the independent review carried out by Professor Dame Wendy Hall and Jérôme Pesenti addresses this issue by providing recommendations to improve access to data. These include a proposal that government and industry should deliver a programme to develop Data Trusts—proven and trusted frameworks and agreements—to ensure exchanges are secure and mutually beneficial. It is also important to be cognisant of the ways in which bias and unfairness operates in the world today, which data reflects.

Establish Accountability Framework

3.9 Though algorithms have long been used, their technical sophistication and breadth of use across sectors is ever increasing. Some applications may be in areas of particular concern, such as public safety or public administration, which deserve increased scrutiny.

3.10 We work hard at Google to ensure our products don't create or perpetuate inequality, however there will not be a panacea that will avoid all risk. When concerns are flagged we act quickly to investigate and resolve them, but also believe a robust impact assessment framework can proactively identify and resolve at least some potential issues. Applying such a framework throughout the value chain of an algorithm supports the goal of accountability, maximising the benefits algorithms afford and mitigating the drawbacks they can present. Google's approach to implementing this is multi-faceted:

- Match the oversight approach with the technical technique. Not all algorithms are created equal, and a given oversight framework may be easier or more effective to apply in some cases rather than others. Decision tree learning, for instance, is a machine learning technique that may lend itself to simpler means of assessing errors than more recent deep learning approaches, albeit at the risk of (sometimes) sacrificing accuracy.

---

[24] https://www.degruyter.com/view/j/popets.2015.1.issue-1/popets-2015-0007/popets-2015-0007.xml
[25] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260
[26] http://www2.datainnovation.org/2014-data-poverty.pdf
[27] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[28] https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

- Consider algorithms in the context in which they are being applied and the potential social and economic impact of their application.
- Scrutinise the data sets used to train the algorithms and seek to ensure that:
    - Data sets are not incomplete, inaccurate, biased, or over or under representative of communities.
    - Data labels are accurate or accounted for—both on the level of individual datapoints, as well as for 'metadata' about the dataset (eg: source, date/manner of preparation, etc)
    - Data sets are appropriate for the use case, with sensitive data used appropriately and in line with applicable laws and the principle of minimising the use of personally identifiable information where possible.
- Internally test and audit the algorithms. While precise procedures will depend on context, this Google research paper "What's your ML test score? A rubric for ML production systems"[29] shares some best practices for the robust testing and monitoring of ML systems.
- Encourage Third party testing without disclosure of source code and data sets. This can help surface unforeseen consequences.
- Engage with civil society. We want to emphasise the benefits of leveraging valuable expertise and experience to identify potentially high risk applications or contexts such as criminal justice and develop solutions. For instance, Google researchers are prominent critics[30] of those attempting to use facial recognition to assess criminality or sexual orientation.

3.11 Algorithms and data can be used to extraordinary effect in spotting and fighting bias and inequality. Indeed, they already are. Google.org partnered with the Geena Davis Institute on Gender in Media to help develop the Geena Davis Inclusion Quotient (GD-IQ), a software tool that uses Google's machine learning technology to help researchers analyze gender representation in popular film with unprecedented speed and accuracy and better equip advocates fighting for gender equity. The Geena Davis Institute recently released "The Reel Truth: Women Aren't Seen or Heard"[31], the first report using findings from the GD-IQ.

3.12 There are great possibilities for the use of algorithms in ways that benefit people, especially underserved communities—and working with companies, organisations, and communities to explore effective examples is a worthwhile goal.

[29] https://research.google.com/pubs/pub45742.html

[30] https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

[31] https://seejane.org/wp-content/uploads/gdiq-reel-truth-women-arent-seen-or-heard-automated-analysis.pdf

- **Whether and how algorithmic decision-making can be conducted in a 'transparent' or 'accountable' way, and the scope for decisions made by an algorithm to be fully understood and challenged;**

3.13 Given the broad range of decision making processes that might be labelled "algorithmic", it is important to recognise that the context in which an algorithm is used is central to determining best practices that should be applied to it. Purposes vary widely: an algorithm might be used to enable someone to find a family photo more easily in their collection, but it also might be leveraged to inform issues of life-changing import, such as housing, education, healthcare, access to finance and issues around criminal justice.

3.14 There are also many ways of configuring the relationship between an algorithm and its users. While it is possible to use algorithms to automate decision-making—and in some instances, such as with spam filtering, this may be the only scaleable approach—it's more often the case that an algorithm will just provide input within a broader human-driven process. This context is hugely important, as it shapes the level of transparency and accountability needed in an algorithmic process.

3.15 These considerations are relevant in the context of leveraging algorithms for decision-making for the public sector. In modern democratic society, there is a general expectation that those responsible for making public policy decisions will strive to make them fairly and appropriately, and that there should be recourse to appeal should concerns arise. Those norms are not changed by technology, but should be incorporated into its design.

3.16 Trust is the key factor. As the use of algorithms in decision-making increases, it is no surprise that society seeks ways to be reassured of the trustworthiness of decisions and to have accountability when these systems fail. Transparency is one means of achieving this end, but it is not the only way. Many technologies in society operate as "black boxes" to the user—microwaves, automobiles, and lights—and are largely trusted and relied upon without the need for the intricacies of these devices to be understood by the user.

3.17 The choice of how best to deliver trust should hinge on the demands of each specific use case, the manner in which algorithms are deployed, and the practicalities of enforcement. Even when transparency is warranted, it is worth noting that there are many ways that it could be implemented. For instance:

- Inputs and Outputs: The functioning of an algorithm might be examined by selectively submitting different types of input and evaluating the outputs in order to provide an indicator of whether it might be producing negative or unfair effects.
- Spotlighting Logic: You could give a visual indication of key metrics relating to an algorithm's functioning without going into the full complexity, akin to the way that car dashboards have gauges for speed, oil pressure, and so on.

3.18 Certain transparency designs may also have their limits. Proposals for code or data to be disclosed in its raw form, for instance, may not be a meaningful or effective means of creating trust or accountability. Just as privacy notices are encouraged to be written in clear, accessible, and easy to understand language in order to provide effective transparency, so too might a flood of technical detail fail to provide adequate notice or understanding about the critical characteristics of a technology.

3.19 These "raw" transparency proposals may also generate their own unique problems. Exposing the code, even if just to a small group in a controlled setting, magnifies the risk of gaming and hacking. There is a real risk that transparency could end up hindering more than helping— bringing more error into the system by making it harder to protect. Search, for example, would be rendered ineffective if the code were made public. We already remove 1 billion spam results every day from spammers seeking to take traffic from more relevant, legitimate websites. The more bad actors know about the search algorithm, the harder it is to protect against such tampering. Moreover, some aspects of algorithms and the data underpinning them will, by nature, be proprietary. Mandating disclosure of such materials would conflict with long-standing data protection obligations and legal protections for trade secrets, as discussed below.

3.20 Google is constantly striving to make the functioning of our products and services understandable to those who wish to know. We do this by explaining what the inputs and outputs are—giving a sense of what data is used, for what purposes—and a high-level description of how the algorithms work.

- It's good for users to understand that our email spam filter relies on data about other messages that have been defined as spam historically, as well as data from security breaches and from customer responses. But precise mathematical details of exactly how the spam classifier works are not necessary (nor feasible to deliver given that the algorithm learns dynamically). Nor are they necessarily desirable, given that exposing how the systems work would make it more likely that spammers would lead to a considerably poorer user experience. What matters is that it works well and it's not doing something against users' intention or wishes.
- For search, we provide a website describing How Search Works[32], over 600 videos on the Webmaster Help YouTube channel[33], and an interactive Search Console tool[34] for webmasters showing errors found on their sites and advising how to fix them, from diagnosing malware to reducing load time.

[32] https://www.google.com/search/howsearchworks/

[33] https://www.youtube.com/user/GoogleWebmasterHelp

[34] https://www.google.com/webmasters/tools/home?hl=en

Research Frontiers

3.21 Algorithms that use modern developments in the field of machine learning present the most difficult challenges in providing effective transparency, due to their complexity. Overcoming the trade-off between interpretability and performance for complex machine learning models is currently among the most researched areas in this space.

3.22 Our ability to understand how such systems render decisions is consistently improving year on year, with work on interpretability happening at major technical conferences, such as NIPS. Advancing this is a priority for Google, not only because it is key to boosting trust in the results of such models, but also because it's likely to yield insights that lead to further improvements. Some examples of Google's projects in this field include:

- Distill: In partnership with Open AI, DeepMind, and others we have established Distill[35], an independent organisation to support a new open science journal and ecosystem supporting human understanding and clarity in machine learning.
- Embedding Projector: We open-sourced an interactive visualisation tool[36] that makes it possible to explore datasets that have hundreds or even thousands of dimensions. This is the same tool that we used internally to help better understand how the neural network behind Translate works.
- Deep Dream: In its earliest incarnation this project was aimed at visualising what different layers within a neural net were learning during training, to make it easier to spot where mistakes in classification arose.
- Glassbox: Glassbox is a machine learning framework optimised for interpretability. It involves creating mathematical models to smooth out the influence of outliers in a data set, thus helping to make results more predictable and decipherable.

3.23 We're optimistic that such efforts will provide us with clearer explanations over time, even if there are limits to what is possible now. While some systems will take more effort than others to 'cut open', with enough resource it's already possible to learn something about how even the most complex models work. For example:

- RankBrain, the machine learning model powering search, is only updated at defined points, allowing us to "freeze" the behavior of the system and make sure it works appropriately before changing it again. Google has a team of search quality raters whose job is to do this auditing, evaluating the usefulness of results per a standard set of relevance guidelines. These ratings don't determine individual page rankings, but are used to help us gather data on the quality of our results and identify areas where we need to improve. These Search Quality Raters use detailed guidelines, which run to over 100 pages, and are entirely transparent and publicly available. Earlier this year, we updated these Search Quality Rater Guidelines[37] to provide more detailed

[35] https://research.googleblog.com/2017/03/distill-supporting-clarity-in-machine.html
[36] https://research.googleblog.com/2016/12/open-sourcing-embedding-projector-tool.html

examples of low-quality webpages for raters to appropriately flag, which can include misleading information, unexpected offensive results, hoaxes and unsupported conspiracy theories. These guidelines will begin to help our algorithms in demoting such low-quality content and help us to make additional improvements over time.

- Google Translate is now powered by a neural network for many languages. To understand more[38] about how this model works, we looked for structure in its mathematical representations of different phrases using the Embedding Projector tool. For instance, we took the same sentence in different languages and compared how they were mapped mathematically, and were reassured to see they were clustered closely if you look at the mapping visually. While this isn't as straightforward as getting explicit rules, it indicates that the system cares more about the semantic meaning of the sentences than it does about what language the sentence is in.

- **The implications of increased transparency in terms of copyright and commercial sensitivity, and protection of an individual's data;**

3.24 This is not a new issue. Certain aspects of algorithms and the data underpinning them will often be proprietary. Mandating disclosure of such information would conflict with long-standing legal protections for trade secrets and, potentially, data protection obligations.

3.25 Currently, UK data protection laws limit data subjects' rights to receive information to that which does not infringe upon trade secrets. Were protections to be watered down by forcing disclosure of such material, in full or part, it could impact the business case for investment in innovation by reducing the rewards for ingenuity because it would undercut the potential scope for new inventions to be a source of differentiation. This would be especially damaging for new entrants, who are likely to find it harder to compete with established players with the resources to swiftly imitate and create a competing service.

3.26 It is an open question whether there are ways to boost transparency without undermining this protection. For some kinds of algorithms, it may be possible to highlight the most influential factors without giving an exhaustive list or details of weighting. In other instances, however, even just revealing an important factor could be unexpectedly damaging. For example, in 1999 Google's founders published a seminal paper on a key innovation in Google's algorithm PageRank, which takes into account the links between pages to assess the importance of websites. Unfortunately, publishing openly about it inspired spammers to game our Search algorithm by paying each other[39] for links, which undermined the factor's effectiveness.

---

[37] The most up to date version of the guidelines can be found here: https://static.googleusercontent.com/media/www.google.com/en//insidesearch/howsearchworks/assets/searchqualityevaluatorguidelines.pdf

[38] https://arxiv.org/abs/1611.04558

[39] https://support.google.com/webmasters/answer/66356?hl=en&ref_topic=6001971

3.27 There is also a wider privacy concern relating to transparency of data. Typically, in order to review the output of an algorithm, you need to know the data that is used as input. However, in some instances this data will consist of personal information. While it's possible to de-identify and aggregate data and limit privacy exposure, that in turn will limit the extent to which you can analyse individual applications of the algorithm, or check for bias or incomplete data.

4. Methods for providing regulatory oversight of algorithmic decision-making, such as the rights described in the EU General Data Protection Regulation 2016.

4.1 The GDPR requires data controllers to provide data subjects with key information on the processing of their personal data. This includes providing meaningful information about the logic involved in, the significance of, and potential consequences of automated decision-making when such decision produces a legal effect on the individual or similarly significantly affects them, or when the decision is based on sensitive data. What precisely is meant by "meaningful information about the logic involved" remains an open question. However, it seems reasonable that this should be interpreted as general, easily understandable information about the types of data used and the functionality of the decision-making algorithm.

4.2 The GDPR also gives data subjects the right not to be subject to a decision based solely on automated processing. The Regulation provides exceptions to this, where such automated decision-making is a contractual necessity, where it is expressly authorised by EU or Member State law, or where the individual has provided explicit consent. In those cases, the individual's rights must be safeguarded by "suitable measures", such as a right to obtain human intervention or to contest the decision.

4.3 Regulatory oversight of compliance with these requirements should therefore involve assessing the level of transparency provided to individuals about algorithmic decisions, and assessing the mechanisms provided for individuals to refuse automated processing or exercise their rights to other safeguards, like requesting human intervention. This will require companies to demonstrate that they have put in place adequate processes to provide accountability for algorithmic decision-making.

4.4 At the moment, such oversight can already be carried out in the same way that regulators enforce other disclosure obligations and data subject rights under the Data Protection Directive. They should be able to easily verify that:

- The data controller's transparency notices include a disclosure of the existence of automated processing where it takes place;
- The controller has provided a description of the automated processing, in a way that makes it intelligible to the individual concerned; and
- In those cases where it is required by the GDPR, the controller has provided individuals with meaningful and easily accessible avenues of redress.

4.5 The development of AI and ML, like any new technology, raises important issues of public trust, which the development of best practices could help to address. Such best practices could also help to ensure that regulatory oversight can rely on a rich set of information.

- For example, companies that regularly carry out automated decisions that have a serious impact on data subjects (not things like spam detection or advertising, but things such as criminal sentencing or approval of a loan) could envisage putting in place internal mechanisms for closer scrutiny of the inputs and outputs and to monitor any unexpected harms.
    - Such a review could focus on the quality of inputs (ensuring that the data fed into the algorithm is not biased, or if it is sensitive data, ensuring that it is adequately protected) and of the outputs (reviewing them to ensure that results are not biased or harmful in expected or unexpected ways).
    - It could be carried out by one of the company's internal functions; or, as appropriate, by a cross-functional group of stakeholders, including external stakeholders, who could bring a diversity of perspectives (technical, ethical, sociological, and other) to the review.
    - These reviews could support regulatory oversight by demonstrating a heightened effort towards transparency as well as the existence of strong internal processes that enhance accountability. The outcome of these reviews could also provide regulators with additional information about the automated processes envisaged, allowing them to more fully assess the accuracy and meaningfulness of the transparency notices provided to users.

- Key stakeholders in AI/ML technology could also help develop better understanding and explainability of algorithms, by funding research on this issue.
    - Such research could span a wide range of disciplines and types of outputs, from developing technical tools that automate explainability, to tools that review automated decisions to scan for any harms or biases, to research that highlights key ethical factors for regulators to focus on when reviewing algorithms.
    - In time, this could help identify more sophisticated best practices, or even seals and certifications, that regulators could rely on for oversight and that individuals could rely on as a measure of transparency and trust.
    - The Partnership on AI is one forum through which such research can be encouraged, but not the only one.

October 2017