Evidence Summary and Core Recommendations

- The responsible use of advanced algorithmic systems brings complex social and technical challenges. The use of algorithms to make decisions where stakes are high, such as in the public sector or highly regulated industries, needs substantially different governance than lower-stakes decisions, such as online recommendations. The Government needs greater capacity to ensure its decision-making algorithms are fair, reliable and accountable, particularly when procured from the private sector.

- Recommendation 1. The Government must ensure adequate and well-trained in-house capacity to develop and understand models, both where algorithms are developed internally, and when they are procured.

- Datasets that power public algorithms are usually created as a by-product of public service delivery, like taxation, justice or education, not gathered specifically for modelling like a survey might be. This data’s meaning, origin, or representativeness are often unclear. For example, we cannot build algorithms based on who breaks the law — only who gets caught.

- Many hands make algorithms, so ultimate responsibility is difficult to assign. Data collection and processing, plus algorithmic design, maintenance, and how the public respond all affect final outcomes.

- Recommendation 2. The Government should engage with universities to study the algorithmic systems they are implementing, the consequences of new systems being designed, and to experiment with new social and technical means of ensuring public values are upheld in real-world deployments occurring today.

- Algorithms are not, and cannot be, ‘neutral’ or independent of the society that produced them. Society must be actively involved in shaping which values are at play in the use of algorithms. It might be logistically easier to hold decision-making algorithms (e.g. determining welfare eligibility) to account than teams of assessors, (with their own subtle biases), but this scrutiny is not easy nor automatic.

- Datasets commonly contain human biases or prejudices, so algorithms powered by such datasets perpetuate these. If we can spot these biases, and agree on a mathematical definition of fairness, we can sometimes statistically mitigate them. Most data scientists lack these skills, so it is rarely done.

- Bias can be complex. Algorithms we train to ‘read’ learn from things humans write, so they pick up many undesirable associations, such as between genders and occupations. Even something difficult to explain, like reading, contains implicit value judgements that we need to actively try to tease out and debate.

- Recommendation 3. The Government should consider developing and mandating algorithm-specific Equality Impact Assessments in order to meet the Public Sector Equality Duty and build public trust.

- Recommendation 4. The Government should integrate transparency and fairness requirements into its IT procurement requirements at all levels to catalyse a world-leading UK market in this sector. This could initially be piloted through systems such as the G-Cloud.

- Transparency does not mean that diverse sections of society can meaningfully hold algorithms to account. Researchers are building technical ways to ‘explain’ the innards of some complex ‘black box’ algorithms, but this probably does not solve accountability on the ground. Neither does forthcoming data protection law, which does not contain a new ‘right to explanation’, contrary to media reports.

- Recommendation 5. The Government should undertake and publish (in line with Recommendation 4 of the MacPherson Review’) an inventory of algorithmic decision-making and decision-support tools at national and local level, and develop metadata standards for publishing their functioning, core logics, performance and remaining uncertainties. As recommended by the Aqua Book, this metadata should be audited for particularly complex and consequential algorithms.

- Recommendation 6. The Government should consider strengthening the General Data Protection Regulation to enable a meaningful ‘right to explanation’ for semi-automated decision-making before it comes into force in May 2018; one opportune moment might be in the debates around the Great Repeal Bill.

- Recommendation 7. To complement this ‘right to explanation’, the Government should invest in interdisciplinary research around how to achieve meaningful algorithmic transparency and accountability from social and technical perspectives.
Introduction to evidence

1. University College London (UCL) welcomes this timely inquiry, which touches upon many areas of our expertise. Much cross-disciplinary activity in this space happening at both UCL and more widely. This report has benefitted from the input of 27 UCL researchers across a range of institutes and departments, who welcome any requests from the Committee for expansion or clarification.

On Terms of Reference Point 1: 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, we believe greater use of data in decision-making should be welcomed, but like all good evidence use, evaluated critically and contextually.

2. Most quantitative understanding of phenomena, or their prediction, is done with algorithms. Technically, any set of automated rules that compute something is an algorithm. This broad definition would include many climate models, economic forecasts, or energy systems models. The 2013 MacPherson Review led to a government inventory of the quality assurance of these type of models, serving to inform the civil service Aqua Book, which serves as useful broad guidance in this space. Many of the recommendations the Aqua Book makes would be as applicable and as beneficial to the algorithmic systems discussed here as to more general government modelling activities. The emerging Government Data Science Ethical Framework is also useful in this regard.

3. The algorithms we see most relevant here are largely predictive machine learning systems designed to anticipate future changes or developments (such as burglaries or firm insolvency) or to detect data that are unobserved (such as whether a tax return is fraudulent or not). Most machine learning systems work by looking at how similar a new case, such as a suspicious credit card transaction, is to old cases, such as a transaction that actually turned out to be fraudulent.

4. While some algorithms make decisions automatically, most of consequence currently only support human decisions. Algorithmic decision-making or decision support has the potential to lead to more informed, efficient decision making, and even more transparent decision-making processes — but this is neither assured nor automatic, and we have much to do to ensure that outcome.

On Terms of Reference Point 2a: The scope for algorithmic decision-making to eliminate, introduce or amplify biases or discrimination, and how any such bias can be detected and overcome, we note that algorithmic bias is a broad, challenging, interdisciplinary scientific field, with many ways of defining ‘fairness’ and barriers to implementing it.

5. Automated decisions adhere to complicated rules that are far from easy to Analyse, and are susceptible to the quality of any input data. As these algorithms are statistical systems, if we believe we have 97% accuracy, we should expect the system to fail 3% of the time. Thus, being clear and open about uncertainties is key, and no decision should be immune from challenge.

6. Machine learning systems are designed to discriminate — to tell the difference — between things like people, images or document, yet there are some types of discrimination that are considered socially undesirable. Some patterns should not be replicated, as they might be related to legal concepts of discrimination, such as avoiding the direct (or indirect) use of protected characteristics, such as gender, disability or so-called race. Others might be to avoid entrenching or retrenching geographic or social inequalities, as these might lead to policy failure. Sometimes, unfairness is less straightforward, such as when we try to teach an algorithm what words mean by feeding it news articles,

7. Indirect discrimination, where variables we did not think were sensitive proxy sensitive ones like race or gender, is a big challenge. For example, imagining a credit scoring algorithm, spending patterns might predict gender (e.g. certain shops), so if gender was predictively useful in some way, a powerful machine learning algorithm might learn to use spending patterns as a proxy.
8. **Algorithms are not, and cannot be, ‘neutral’ or independent of the society that produced them, and as such society needs to be actively involved in shaping which values are at play.** Most algorithms learn from data collected in society, and this data cannot be made independent of the society. It also cannot be made independent from its designers. According to a common definition, we say a machine learns from experience with respect to some class of tasks and performance measure, if its performance at the tasks, as measured by the performance measure, improves with experience. Designers choose what counts as valid experience, an appropriate task, and an idea of performance that tries to capture ‘good’ and ‘bad’. There is no way to neutrally ‘optimise’ these value-laden choices, and decisions at this stage often have unintended consequences downstream.

9. **Some researchers have been attempting to correct algorithms by mathematically formalising political and legal notions of discrimination and inequality.** These notions include demographic parity (women and men always treated equally); equality of opportunity (focusing on non-discrimination only in those who don’t get a preferred outcome, like being granted a loan); or ‘counterfactual fairness’, where we try to estimate what would have happened had a protected characteristic, such as gender, been different, given all the implications that difference might bring. To remove the effects of any given traits from these systems, you might massage the datasets, change the way the machine learning occurs, or change the final algorithm after it has been created. This is general is however very hard, as variables other than the protected characteristics might give information about them, or certain subsets of the data might contain discrimination even if the overall data does not.

10. **Mathematical notions of fairness are often not straightforward to discuss and debate.** While the public might often want some or all of these notions of fairness, many of them have been proven to be mathematically incompatible with each other. Additionally, mathematically removing the effect of a feature in a system does not automatically lead to public trust. Technology engagement tools, such as those from the field of Responsible Research and Innovation, might be useful to build on here.

11. **In many common algorithmic areas, such as asking computers to interpret text (Natural Language Processing), what is ‘biased’ or ‘fair’ is difficult to talk about technically as well as socially.** For example, some of the most popular technologies to enable machines to ‘understand’ text turn each word into about 200–500 numbers, generated from running neural networks on very large datasets of text, such as newspaper articles from the last 10 years, or all of Wikipedia. By comparing these numbers, we can see that ‘man’ is to ‘woman’, as ‘king’ is to ‘queen’. Unfortunately, by this logic, and due to the biases in real world text needed to train, ‘man’ is to ‘woman’ as ‘computer programmer’ is to ‘homemaker’. Some researchers have tried to identify such implicit biases, but they are difficult to define — and the sheer number of problematic relations we might find in these kind of algorithms makes it hard to know which to prioritise, or when to stop.

12. **In addition, these theoretical debiasing approaches would not be straightforward to implement in practice due to practical and regulatory constraints.** Current statistical approaches to debiasing algorithms require access to sensitive characteristics, yet on the ground we often need to know potential issues before we collect or link data (for example, to fill in a Privacy Impact Assessment), or have obligations under data protection law to minimise the data we collect, particularly sensitive data. Researchers considering these issues have outlined stepping-stone approaches, such as third party auditors empowered to collect sensitive, data or a crowdsourced knowledge base recording fairness issues across sectors, datasets and applications.

13. **Public bodies, through the Public Sector Equality Duty (PSED), also have a positive obligation to reduce discrimination and actively promote equality, which needs technical translation into analytics and machine learning systems.** The PSED requires substantive, and rigorous consideration of these issues before policy deployment, and on a continuing basis thereafter (the Brown Principles). This might involve inserting new logics or positive discrimination into systems, which has some precedent but needs collaborative research and active practice development. It is not clear the PSED is being substantively followed without these considerations, particularly as Equality Impact Assessments are no longer mandatory.
14. **Bias can also emerge from broader processes and contexts, such as data collection, transformation, modelling and decision support, and the Government must work with researchers in real-world contexts to understand this.** Ongoing UCL research around the UK and four other governments’ uses of machine learning today has been illuminating how debiasing methods, while useful, are unlikely to succeed in diagnosing and mitigating complex issues alone.\(^{16}\) Examples of successful collaboration to understand these phenomena exist — some of the earliest research explicitly concerning fairness in public algorithms, in 2009–10, has been undertaken with the Dutch Ministry of Justice and the Dutch Police\(^ {17}\), while UCL researchers have been engaged with how predictive maps for aiding police forces are used on the ground since 2007.\(^ {18}\) The Jill Dando Institute Research Laboratory at UCL, the first Police Assured Secure Facility for research data in Europe, is exceptionally well placed to support UK researchers to analyse real data.

15. **Machine learning situations have been developed in low-stakes private sector environments, like online shopping.** Public sector applications differ strongly, so bring new challenges for research and practice. Public sector organisations are primarily interested in modelling rare, high stakes events, such as child abuse, burglaries, tax or benefit fraud, and so on. This can be different from the logic behind machine learning systems on the web, where information display is common, low-stakes and comparatively costless. Similarly, decisions made by influential bodies like governments or the police can change the data we collect tomorrow — for example, as criminals try to ‘game’ the system\(^ {19}\), or as self-driving cars and society adapt to each other\(^ {20}\). ‘Black box’ machine learning systems are much more difficult to adjust for these feedback effects than theoretically grounded systems, like crime models based on theories of ‘contagion’ across neighbouring areas.\(^ {21}\)

16. **Machine learning algorithms have aims, assumptions and values ‘baked in’ by virtue of their design, and these need to be challenged and discussed depending on the context.** In order to train algorithms, we have to tell them what we want them to do and not to do.\(^ {22}\) For example, we sometimes want to avoid false negatives (e.g. an automated car not stopping when it should), while other times we want to avoid false positives (e.g. investigating an innocent family for child abuse). This can be more context and problem specific. For example, sometimes we want to consider not just how good our algorithm is at predicting crime in particular neighbourhoods, but whether it gives us ‘patrollable’ results, rather than small dispersed patches all over the place.\(^ {23}\) This means we have to engage well with end-users, such as police officers, when designing any systems, and that ‘off-the-shelf’ algorithms might be inappropriate.

17. **Focussing too much on improving operational decision support might mask the fact that your policy approach is wrong, and you might need a different intervention.** If the type of disability allowances you provide is not suitable for certain segments of society, then no amount of decision support will ‘clean’ this bias. If tax fraud investigators are not skilled enough to investigate certain types of fraud when it is present, then better directing them to it with algorithms is unlikely to provide useful policy support — better investment in training might be more appropriate.

18. **Recommendation 1. The Government must ensure adequate and well-trained in-house capacity to develop and understand models, both where algorithms are developed internally, and when they are procured.**

19. **Recommendation 2. The Government should engage with universities to study the algorithmic systems they are implementing, the consequences of new systems being designed, and to experiment with new social and technical means of ensuring public values are upheld in real-world deployments occurring today.**

20. **Recommendation 3. The Government should consider developing and mandating algorithm-specific Equality Impact Assessments in order to meet the Public Sector Equality Duty and build public trust.**
On Terms of Reference Point 2b–2c: 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; and the implications of increased transparency in terms of copyright and commercial sensitivity, and protection of an individual’s data, we note that systems are not necessarily ‘black boxes’, and algorithmic transparency, while often important, is not the same as accountability.

21. Well-documented and accessible algorithms might provide insight into public decision-making, and increase accountability. A central register of these technologies, with meaningful open code and/or metadata about them, would improve democratic process. A central tension with making algorithms completely open is that many are trained on personal data, and some of this private data might be discoverable if we release the algorithmic models. This needs to be managed carefully.

22. Where algorithms are used to manage complex systems, it is usually technically possible to make much more open, accountable, transparent, and/or privacy preserving systems than we are deploying today. Many researchers favour decentralised systems, where there is no single place where data is stored or verified, such as a blockchain. These provide many perceived benefits, such as independence from political control and public verifiability of activity, and security against certain types of interference or attack. UCL research has illustrated ways in which central government and decentralised technologies can work together to get the best out of difficult trade-offs. For example, raw data from smart metering technologies can be extremely privacy-invading. Direct access to smart meter energy reading might tell someone not just whether you are at home, but what television channel you are watching, or webpages you are viewing. Yet technologies utilising encryption in innovative ways exist to undertake all the purposes of smart meters (e.g. billing, load monitoring, efficiency analysis, fraud detection etc.) without compromising individual data records. Despite this, as large incumbents, such as energy firms, are usually unaware of these options, and the form of proposed infrastructure is usually set up before, not during tendering processes, there are high barriers to the adoption of world-leading technologies.

23. In other cases, particularly around machine learning, there are technologies being developed to ‘open’ black-boxes. A range of technologies have been proposed to extract rules and core logics from these technologies in a way that allows them to be communicated. While this might not always be easy, technologies such as these should be integrated where possible into the practice of public machine learning. At the moment, there are insufficient legal requirements for this (see para 26).

24. Recommendation 4. The Government should integrate transparency and fairness requirements into its IT procurement requirements at all levels to catalyse a world-leading UK market in this sector. This could initially be piloted through systems such as the G-Cloud.

25. Recommendation 5. The Government should undertake and publish (in line with Recommendation 4 of the MacPherson Review) an inventory of algorithmic decision-making and decision-support tools at national and local level, and develop metadata standards for publishing their functioning, core logics, performance and remaining uncertainties. As recommended by the Aqua Book, this metadata should be audited for particularly complex and consequential algorithms.

On Terms of Reference Point 3: Methods for providing regulatory oversight of algorithmic decision-making, such as the rights described in the EU General Data Protection Regulation (GDPR) 2016, we note that the text of the GDPR does not effectively provide the rights that the media have attributed to it, and that legal remedies might not be appropriate for all cases.

26. Data protection law offers individuals some useful rights, yet because many significant algorithmically-driven decisions are not fully automated, citizens cannot draw on them when they would matter most. As Turing Institute researcher Sandra Wachter and colleagues note, the text of the General Data Protection Regulation (GDPR) does not provide the same right to explanation that is promised in the recital, which is legally problematic. A widely circulated but misleading short conference paper has led many in government, industry and media to believe a
new ‘right to explanation’ will soon exist.\textsuperscript{30} Individuals cannot ask for explanations of algorithmic decisions where a ‘human-in-the-loop’ exists, even though much research notes that humans in the loop often over-rely on decision support systems, and so what counts as \textit{de facto} ‘automated’ is unclear.\textsuperscript{31}

27. **There is also a large practical gap between having a technically or legally transparent system, and enabling citizens to meaningfully understand decisions that concern them.** The number of algorithmic decision support and decision-making systems is growing, but the cost of data scientists and the squeezed training pipeline has meant that effective oversight by third party organisations and civil society is not straightforward to achieve. Empowered, trusted brokers are needed to communicate the state-of-the-art in technologies, applications and concerns to and from a range of societal actors, without individuals having to seek costly, risky and time-consuming legal remedies.

28. **Recommendation 6. The Government should consider strengthening the General Data Protection Regulation to enable a meaningful ‘right to explanation’ for semi-automated decision-making before it comes into force in May 2018; one opportune moment might be in the debates around the Great Repeal Bill.**

29. **Recommendation 7. To complement this ‘right to explanation’, the Government should invest in interdisciplinary research around how to achieve meaningful algorithmic transparency and accountability from social and technical perspectives.**

April 2017

**Acknowledgements**

The lead author of this submission was Michael Veale, doctoral researcher in responsible public sector machine learning at the Department of Science, Technology, Engineering and Public Policy (STEAopp). The comments, text, time and feedback of the following members of the UCL community greatly helped shape this document: Sofia Olhede, Patrick Wolfe (Department of Statistical Science, UCL Big Data Institute); Christian Hennig, Ricardo Silva (Department of Statistical Science); Jon Agar, Phyllis Illari, Melanie Smallman, Tiago Mata, Jack Silgoe (Department of Science and Technology Studies); Kate Bowers, Shane Johnson, Toby Davies (Department of Security and Crime Science, see their separate evidence in addition); George Danezis, Zeynep Engin, Sarah Meiklejohn, Robert E Smith, Johannes Welbl (Department of Computer Science); David Tuckett, Sam Stern (Centre for the Study of Decision-Making Uncertainty); Sarah Chaytor, Graeme Reid (UCL Public Policy); Chris Tyler (UCL STEaPP); Louise Chisholm (UCL eResearch Domain); Richard Moorhead, Sylvie Delacroix (UCL Laws); Rose Luckin (UCL Knowledge Lab). While this document represents a range of views sought from across the College, it does not state UCL opinion.

\begin{references}
\item HM Treasury (2013). ibid.
\item Machine learning systems are algorithms that identify and utilise patterns in data, usually for prediction.
\item The Committee should be aware that in statistics, \textit{bias} has a specific and distinct meaning in relation to the evaluation of models against the data you are trying to model.
\item The Economist (11 Apr 2013). How might your choice of browser affect your job prospects?
\item See the UCL RRI-TOOLS project: The RRI Toolkit. http://rri-tools.eu
\item Veale, M and Binns, R [under review]. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data.
\item Rocktäschel, T et al. (2015). Injecting Logical Background Knowledge into Embeddings for Relation Extraction. HLT-NAACL; Verwer, S; Calders, T (2012) Introducing Positive Discrimination in Predictive Models. 10.1007/978-3-642-30487-3_14
\item This is an issue currently under active research by Michael Veale and Irina Brass at UCL STEaPP.
\item Veale, M. Ongoing PhD research between UCL STEaPP and GO Science. [documents on request]
\item Johnson, S et al. (2007). Prospective crime mapping in operational context. Home Office.
\item Veale, M. Ongoing PhD research between UCL STEaPP and GO Science. [documents on request]
\end{references}
21 See separate submitted evidence by Kate Bowers, Shane Johnson, Toby Davies and colleagues at UCL Security and Crime Science, and work on feedback issues around police use of machine learning today by Michael Veale at UCL STEaPP.
22 These instructions are called loss functions. See Hennig, C; Kutlukaya, L (2007) Some thoughts about the design of loss functions. REVSTAT 5, 19–39.