I. Introduction
The Oxford Internet Institute makes this submission to the Committee as part of the inquiry on “algorithms in decision-making” as lodged by the House of Commons’ Science and Technology Committee on 28 February 2017.

The Oxford Internet Institute is honoured to see that the Committee lodged this inquiry based on the work of its researchers Dr. Sandra Wachter, Dr. Brent Mittelstadt, and Prof. Luciano Floridi “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation”\(^1\) and welcomes the opportunity to provide evidence.

The Oxford Internet Institute is a multidisciplinary research and teaching department of the University of Oxford, dedicated to the social science of the Internet. Our academic faculty and graduate students are drawn from many different disciplines: we believe this combined approach is essential to tackle society’s ‘big questions’. Together, we aim to positively shape the development of our digital world for the public good.

II. Executive Summary
The Oxford Internet Institute

1. calls for a legally mandated “right to explanation” of automated decisions to ensure that decisions are fair and transparent and can be successfully contested;

2. calls to establish an algorithmic oversight institution, such as an independent or regulatory body, and equip it with sufficient expertise, resources, and auditing powers. Such a third party would strike a balance between the legitimate interest of people to understand how and why an algorithm reached a specific decision about them and the interest of data controllers, such as companies, to protect their trade secrets and other vital interests;

3. calls to establish auditing and certification mechanisms (such as “seals”) for algorithms, which will test algorithms for unethical consequences prior their deployment, as well as identify algorithms whose design, development, and deployment produces unbiased, fair, non-discriminatory, and accurate outcomes. This approach would also prevent data controllers from disclosing their trade secrets;

4. calls to mandate the use of human interpretable algorithmic decision-making methods in highly sensitive or risky sectors, such as insurance, healthcare, education, the workplace, judicial sentencing and predictive policing;

5. calls for investment in further research into the ethical, legal, and technical feasibility of explanations of algorithmic decisions, algorithmic “auditing” procedures, and unbiased design of algorithms.

III. Inquiry

a. 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

Automated decision-making is widely applied across various sectors, such as insurance, healthcare, workplace management, human resources, and criminal justice systems. Increasingly, machines assess and categorise human beings. Automated decision-making in health insurance is one critical area. Algorithms are deciding on the insurance eligibility of applicants and undertaking risk assessments to determine premiums. The same holds true for AI used for hiring, firing, and promotions. Algorithms are also used by judges in making decisions about sentencing and parole. These systems calculate the probability of a person to reoffend. Predictive policing is also a very sensitive area where the police use these technologies to identify criminal hotspots and individuals at risk of offending. Automated systems used in warfare raise similar concerns regarding the identification of and response to threats.

Healthcare, employment, and criminal justice are areas with significant impact on quality of life. When a decision is taken about individuals in these areas, it is natural to want to make sure the decision is fair—that people are not discriminated against or treated unfairly. This holds true across automated decision-making in government and commercial contexts. The usage of algorithms to make decisions is itself not a problem, for it may help make fairer and more accurate decisions. It may prove to be the case, for example, that algorithms can make less biased decisions than humans if programmed and trained well or given high quality, unbiased data. Algorithmic decision-making only becomes a problem when we do not know how a given algorithm reached its conclusion and how the actions based on these conclusions adversely affect us, such as being fired, denied parole, or declined for insurance. We have a tendency to trust these technologies more than human assessment, simply because they consider lots of data. We view them as ‘smart’ or ‘intelligent’, at times even ‘objective’, even though they are often opaque ‘black boxes’ that are difficult to scrutinise. Clearly, before any trust is placed in complex, learning algorithms, appropriate accountability mechanisms need to be in place. Following this, a “right to a meaningful explanation” should be granted to individuals. Such a right is not currently guaranteed by the General Data Protection Regulation (GDPR), as explained below.

b. Whether ‘good practice’ in algorithmic decision-making can be identified and spread

b.1. The scope for algorithmic decision-making to eliminate, introduce or amplify biases or discrimination, and how any such bias can be detected and overcome.

Human decision-making is inevitably biased. Decision-making algorithms are often thought to be free from the innate biases and blind spots of human decision-makers, and less prone to make unfair or discriminatory decisions. However, it is difficult to
prove this assertion without meaningful oversight of the complex, often opaque decision-making algorithms and the datasets that they consider to make decisions. Datasets often reflect latent social and technological biases; bias and discrimination may emerge only when an algorithm processes particular data. Sustained oversight of algorithmic decision-making processes is required.

Algorithmic decision-making can also harm individuals in new and unique ways compared to human decision-making, for which existing ethical and legal frameworks struggle to provide a solution. This is the case of group discrimination (e.g. ageism, ethnicism, sexism), for example. Group discrimination occur when individuals are discriminated not because of their identity but because they match a specific ‘type’ identified by an algorithm (e.g. the cat-owner, the pizza-eater, or part of the LGBT+ community), independently of the identification of each member of the group. Because of their focus on individuals, current ethical frameworks, data protection, and anti-discrimination law only partially address the challenges of algorithmic decision-making.

Automated decisions can replicate harmful and discriminatory aspects of human decision-making and potentially create new types of harms. Further work is therefore required to identify and foresee new types of harm, and to develop methods that automatically detect and, ideally, rectify problematic decisions as they occur. Without further development of these areas, affected individuals and groups have little prospect of rectifying wrongs when they occur.

We believe that significant support needs to be offered to develop methods and standards for detecting and explaining harms in algorithmic decision-making. Without this essential work, algorithmic decision-making cannot reasonably be treated as accountable, transparent, or fair. Work on algorithmic auditing is one such area requiring much greater attention from ethicists, legal scholars, data scientists, policy-makers, and regulatory bodies. Current approaches to auditing the functionality and impact of decision-making algorithms are largely based on social scientific audit studies. Further development is needed on auditing technologies and ethical frameworks to guide them, including alert functions embedded in algorithmic systems and external auditing algorithms. Auditing can ideally create a procedural record to demonstrate bias against particular individuals and groups. It can help data controllers to meet accountability requirements by detecting when decisions harm individuals and groups, by explaining how they occurred, and under what conditions they may occur again. However, auditing is not feasible without strong regulatory support and cooperation from data controllers.

b.2. 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.

This question needs to be understood from an ethical, legal, and technical view. While technological research and development currently focuses on interpretable methods of algorithmic decision-making, very little research is being done on ethical and legal dimensions of transparency and accountability. Current legislation (see
Written evidence submitted by the Oxford Internet Institute (ALG0031)

below) does very little to protect individuals from discrimination when automated decision-making is used. Clearly, this is a serious problem. Solving it requires a careful consideration of the ethical problems posed by automated decision-making to determine the principles and values that we want to protect. This is a preliminary and necessary step to the definition of appropriate regulatory mechanisms. A discussion is required of how to implement these principles in law, policy, guidelines, and design. This discussion needs to connect ethical and legal views with technical expertise to ensure that the desired outcomes are not just aspirational but practically feasible. As an example, it would be important to consider laws that would require human interpretable systems[1] in sensitive areas, such as the health sector, employment, or predictive policing.

It is pivotal that further research is funded to answer these questions. Additionally, if we arrive at the conclusion that explanations are not technically feasible, we need to explore other tools, such as auditing mechanisms (see above) to prevent and detect discrimination and biased outcomes.

b.3. The implications of increased transparency in terms of copyright and commercial sensitivity, and protection of an individual’s data.

The research that inspired this inquiry revealed that, according to current jurisprudence and legal commentary in Europe, two of the biggest hurdles to a “right to explanation” (see below) are trade secrets and copyright. When people have asked for an explanation of an individual decision in the past, data controllers have been reluctant to share information about the decision-making process, due to fears that this could hurt their business interests. Legal commentators have acknowledged the legitimacy of business interests, and likewise suggest that the General Data Protection Regulation (GDPR) and the (old) European Data Protection Directive do not offer sufficient grounding for a “right to explanation”.

In the same vein, the aforementioned research revealed that data controllers have thus far been required to reveal only limited information when algorithmic decisions have been legally challenged. Under the Directive, data controllers have generally not been required to reveal particular details about the algorithms used, including code, statistical values, criteria, reference or comparison groups, and weighting of elements to calculate probabilities (e.g. the likelihood of loan repayment).

The GDPR is likely to follow the precedent set by the Directive, at least initially. This would be unfortunate. The rapid spread of automated decision-making into sensitive areas of life, such as health insurance, credit scoring or recruiting, demands that we do better in allowing people to understand how their lives are being shaped by algorithms.

Even though it is understandable that data controllers have a legitimate interest not to disclose trade secrets, this interest does not always outweigh the interests of individuals to understand how algorithms make decision about them. In order to balance these competing interests, we propose an AI Watchdog[2], or a trusted and independent regulatory body. This body would need to be equipped with
Written evidence submitted by the Oxford Internet Institute (ALG0031)

the proper expertise (spanning ideally law, ethics, to computer science), resources and auditing authority (to make inspections) to ensure that algorithmic decision-making is fair, unbiased, and transparent.

We further support the establishment of certification mechanisms for decision-making algorithms. Certification mechanisms, such as seals, can likewise guarantee that algorithms are designed with fairness and accountability in mind, and do not require trade secrets to be publicly disclosed.

c. Methods for providing regulatory oversight of algorithmic decision-making, such as the rights described in the EU General Data Protection Regulation 2016
Since the approval of the GDPR in 2016, it has been widely and repeatedly claimed that this framework will legally mandate a ‘right to explanation’ of decisions made by automated or artificially intelligent algorithmic systems. This right to explanation is viewed as an ideal mechanism to enhance the accountability and transparency of automated decision-making. However, the ‘right to explanation’, as described is in the GDPR, is in reality more a ‘right to be informed’. Although it appeared in earlier drafts[3] in the trilogue negotiations, the “right to explanation” is only mentioned in Recital 71, which is a legally not binding provision, as Recitals only give guidance on how to interpret the law[4]. The legally binding Article 22 that focuses on automated decision-making does not include this right.

The other sections that deal with automated decision-making, Articles 13-14 (notification duties) and Article 15 (the right of access), also fail to provide individuals with a right to an explanation. These Articles state that individuals have the right to receive information about “the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.” Based on legal commentary and jurisprudence, data controllers only need to inform individuals about the usage and basic design of algorithmic decision-making methods, but not about the details or steps they took to make a specific decision.

Following how comparable requirements in the (old) European Data Protection Directive were implemented by Member States and interpreted by national courts, the information that individuals are entitled to receive concerning automated decision-making will likely be heavily limited by trade secrets and other interests of data controllers. There is, therefore, a risk that the GDPR will entitle individuals to very little meaningful information about automated decisions.

In addition, the safeguards (the right to obtain human intervention, to express his or her point of view, and to contest the decision) in Article 22 relating to automated decision-making have very limited applicability. Safeguards are only required when the algorithmic processes that are used are solely automated. Consequently, the safeguards do not apply if humans are put in or on the decision-making loop at any point, even if they do not plan to intervene in the decision-making process in any meaningful way. Further, the safeguards only apply when the outcome of the decision is considered to have ‘legal or other significant effects’ on an
individual. The meaning of ‘legal’ and ‘significant’ effects, and thus the scope of the Article 22 safeguards, remains unclear.

Even when these safeguards apply, it is questionable whether they are practically useful without a complementary right to explanation. Challenging a decision, for example, is difficult if the person does not know how it was reached. The right to obtain human intervention and to express one’s views will likewise not be helpful if the person concerned does not understand how the decision was made.

Seeing that the GDPR lacks precise language, future legislative steps should close these and other gaps identified in the aforementioned foundational research. This leads us to conclude that a legally binding “right to explanation” should be established.

This right is crucial. We must understand how these systems reach their conclusions, so we can contest them, assess their accuracy, and hold the organisations using them accountable.

April 2017

[1] An example of legislation already requiring interpretable methods in automated data processing is the US Fair Credit Reporting Act.