Before the
House of Lords
Select Committee on Artificial Intelligence
Comments of Electronic Frontier Foundation
September 6, 2017

Peter Eckersley, Ph.D.; Jeremy Gillula, Ph.D.; Jamie Williams
Electronic Frontier Foundation
815 Eddy Street
San Francisco, CA 94109
United States of America
Telephone: +1 (415) 436-9333
Email: pde@eff.org

1) The Electronic Frontier Foundation (EFF) submits the following comments in response to the House of Lords Select Committee on Artificial Intelligence's Call for Evidence, available at http://www.parliament.uk/documents/lords-committees/Artificial-Intelligence/Artificial-Intelligence-call-for-evidence.pdf. EFF is a member-supported, nonprofit, public interest organization composed of activists, lawyers, and technologists, all dedicated to protecting privacy, civil liberties, and innovation in the digital age. Founded in 1990, EFF represents tens of thousands of dues-paying members, including consumers, hobbyists, computer programmers, entrepreneurs, students, teachers, and researchers. EFF and its members are united in their commitment to ensuring that new technologies are not used to undermine privacy and security.

2) For the purposes of our comments we use a fairly broad definition of AI, which includes everything from simple machine-learning (ML) systems to advanced deep-learning techniques. While others may focus their remarks on more advanced systems, we believe it is important to acknowledge that even simple AI systems in use today (which some may no longer even classify as AI) are already having a dramatic impact on society.

1. What is the current state of artificial intelligence and what factors have contributed to this? How is it likely to develop over the next 5, 10 and 20 years? What factors, technical or societal, will accelerate or hinder this development?

2. Is the current level of excitement which surrounds artificial intelligence warranted?

3) We have made some initial studies of the pace of technical progress with our AI Progress Measurement initiative (available at https://www.eff.org/ai/metrics), which surveys problems, metrics, and benchmarks from the machine learning research literature, and tracks progress on them.

4) Any prediction of future technological development is prone to significant limitations and methodological difficulties, so we wish to stress that the pace of technical progress over the next 5, 10, and 20 years is of course highly uncertain. However, the data we have collected shows

that this field is making rapid advances on a very wide range of problems. Influxes of talent, resources, and computing power will likely continue this trend.

5) Although there remain many daunting obstacles and difficult tasks that AI is not yet close to solving, there is evidence to support claims that machine learning could have significant economic impacts in a growing number of domains over the next 20 years, and that there is some possibility of drastically transformative AI technologies emerging in the next 10-30 years.

7. How can the data-based monopolies of some large corporations, and the 'winner-takes-all' economies associated with them, be addressed? How can data be managed and safeguarded to ensure it contributes to the public good and a well-functioning economy?

6) This question combines several different issues. Winner-takes-all monopolies are not a new phenomenon in the computing industry, which often exhibits strong economies of scale, lock-in effects and even stronger network externalities, where the usefulness of a product is proportional to the number of people already using it. Many of the present concerns about big data and market power are extensions of these pre-existing and unsolved policy problems in the technology industry. We will not attempt to address them in this submission.

7) But there are also several competition policy questions that are quite specific to machine learning and AI research. These essentially derive from two pre-conditions: (1) present machine learning techniques require an enormous number of examples to successfully learn things; and (2) typically, only large technology companies are in possession of enough examples--i.e., the photos, emails, text messages, location histories, and/or sensor feeds of hundreds of millions of people--necessary to conduct machine learning research. This has given large companies a significant advantage when conducting basic research on certain machine learning problems.

8) Fortunately, this lead is not universal across all machine learning problems. For certain tasks, the academic community has been able to build its own large datasets that are comparable to privately held ones, or at least sufficiently large enough to achieve research breakthroughs. In other cases, technology companies have voluntarily shared data in order to stimulate open research on problems they consider important, and/or to promote themselves to prospective employees.

9) But for other tasks, the process of sharing datasets is somewhat complicated by the private and sensitive nature of the data required. For example, if one wishes to use machine learning to understand and process email better (for instance, in making a better spam filter or in making an agent that can read and handle some types of email for you), one needs large datasets made from genuinely representative sample emails. Since it is inherently problematic to share large, representative datasets of people's private email, only major email providers can directly perform this sort of machine learning research. Similar problems apply to research based on network traffic data, server logs used for cybersecurity purposes, patterns of online behavioural data, and many other categories.

10) In the long run, we are unsure how serious a problem this will be. New algorithmic techniques such as federated learning[1] and differential privacy[2] could in theory allow more sharing of privacy-sensitive training data, but it is doubtful they could fully close the productivity gap between AI researchers working for the largest tech companies and the rest of the research community.

11) Though these effects are likely to continue to confer advantages to established players in the race to apply machine learning to some economically important tasks, it is less clear that there are sound competition policy interventions available in the AI space specifically at this time. Probably the most constructive role that governments could presently play is to provide additional incentives and support for the creation of open research datasets, particularly where there are algorithmic ways to solve the privacy and security problems that would otherwise hamper the use of that data for research purposes.

8. What are the ethical implications of the development and use of artificial intelligence? How can any negative implications be resolved? In this question, you may wish to address issues such as privacy, consent, safety, diversity and the impact on democracy.

12) The ethics, safety, privacy, and social policy questions raised by existing machine learning technologies are quite complicated, and they will grow much more so as artificial intelligence becomes more autonomous and capable of more diverse and learned forms of action. There is an emerging field of academic and industry research focused on these questions, but they are far from solved.

13) In one pressing short-term example, the use of machine learning and other statistical algorithms is already disparately and unfairly impacting different populations in society due to problems that include not only biased training datasets but also the emulation of human prejudices as a result of a statistical problem called *omitted variable bias*. Omitted variable bias occurs when an algorithm lacks sufficient input information to make a truly informed prediction about someone, and learns instead to rely on available but inadequate proxy variables. For instance, if a system was asked to predict a person's future educational achievement, but lacked input information that captured their intelligence, studiousness, persistence, or access to supportive resources, it might learn to use their postal code as a proxy variable for these things. The results would be manifestly unfair to intelligent, studious, persistent people who happened to live in poorer areas.

14) Detecting and analyzing this sort of unfair impact is complicated by the fact that, depending on the context, it is not always clear what the appropriate measure of bias should be--or in other words, what results would be "fair" and what results would be "unfair." The topic of

[1] See B. McMahan et al, Communication-Efficient Learning of Deep Networks from Decentralized Data, AISTATS 2017, https://arxiv.org/abs/1602.05629.

[2] See C. Dwork 2014, Algorithmic Foundations of Differential Privacy https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

“algorithmic fairness” has spawned an entire research field, some aspects of which we discuss in our answer to question 10 below.

15) The current deployment of machine learning techniques poses a serious privacy risk. Specifically, machine learning has enabled efficient large-scale surveillance both by intelligence agencies and commercial actors. In the past, large-scale surveillance of a population was limited by the human resources available to sift through the data collected. Only societies like East Germany, that were willing to recruit one informant per 6.5 citizens, could possibly watch and pay attention to all of their citizens’ actions. But the combination of already-deployed surveillance technologies and machine learning for analysing the data will mean that exhaustive surveillance is becoming possible without the need for such enormous commitments of money and labour. The potential of machine learning to enable such effective large-scale surveillance has reduced the price tag of authoritarianism, and poses a novel threat to free and open societies. For this reason, EFF believes that machine learning algorithms should only get access to a person’s data with their consent and control, or a properly issued warrant.

16) More broadly, when algorithms make decisions that affect human lives in ways that may be mundane but expensive (e.g., price discrimination) or profound (e.g., sentencing or bail recommendations, use of machine learning in mass surveillance), there must be transparency, openness, due legal process, and accountability for intended and unintended consequences, as we describe in our response to the next Question 9.

17) In the medium term, the impact of AI on labour markets deserves serious attention. The EFF has no view on the right solution to that problem, but we do think there is some risk of the market providing many fewer well-remunerated jobs in the coming decades, or many fewer jobs in general, and that planning in advance for this possibility is an important task for societies that are presently fundamentally motivated by and organised around the Protestant work ethic. We would urge those across the ideological spectrum to think seriously about what kinds of society they would want to see if less human labour was practically necessary for prosperity. Would we be comfortable with fewer people working, and willing to share resources with them? Would it be better to create new forms of artificial work? How will we preserve the sense of opportunity, self-worth and status of humans in such societies?

18) In the longer term, the ethical and societal questions around AI may be even stranger and more profound. What would it mean for humanity to share the planet with other types of intelligence? Though entertaining, such topics are extremely speculative and at present more usefully addressed by academic research and science fiction than by concrete policy making, though we do think that there are some exceptions. For instance, if highly transformative artificial intelligence technologies were developed in the future, the risks associated with computer insecurity would almost certainly rise dramatically. As a result, we believe the possibility of AI advances in coming decades are a reason to increase funding and incentives for the creation of secure computing infrastructure and effective defensive cybersecurity systems today.

9. In what situations is a relative lack of transparency in artificial intelligence systems (so-called 'black boxing') acceptable? When should it not be permissible?

19) Any AI systems that significantly impact the rights, freedoms, or lives of large populations of people must at least be auditable, if not transparent. Examples of systems where some level of transparency is necessary include:

- AI systems used for government purposes (e.g., to advise judicial decisions, to help decide what public benefits people do or do not receive, and especially any AI systems used for law enforcement purposes);
- AI systems used by companies to decide which individuals to do business with and how much to charge them (e.g., systems that assign credit scores or other financial risk scores or financial profiles to people, systems that advise insurance companies about the risk associated with a potential customer, and systems that adjust pricing on a per-customer basis based on the traits or behavior of that customer);
- AI systems used by companies to analyze potential employees; and
- AI systems used by large corporations to decide what information to display to users (e.g., search engines, AI systems used to decide what news articles or other items of interest to show someone online--if they make those decisions based on individual user characteristics--and AI systems used to decide what online ads to show someone).

20) The appropriate level of transparency will be different for each of the scenarios described above. For example, given the tremendous impact AI systems used for government purposes or financial decisions can have on people's lives, such AI systems should be completely transparent--regardless of whether or not they were publicly or privately developed. The public and all those potentially impacted should have access to the algorithms (and as we describe below, training data), and the systems should be subject to regular, published audits, which should include measuring how the system performs under various fairness metrics, to ensure that the system continues to function as expected (and is not causing any discriminatory, unfair, or unintended effects). These audits could be performed by the organization responsible for the AI system or an independent governmental body, but they must be mandated to ensure that they are performed in a regular and timely manner. And of course, all audit results should be immediately made public.

21) A lower level of transparency may be appropriate for algorithms that have a lesser impact on people's lives, such as search engines and news feed algorithms (although the impact of these algorithms is by no means insubstantial). These algorithms are often closely guarded trade secrets that required tremendous R&D expenditure to get right. As such, a lower level of transparency, such as effective auditability for discriminatory outcomes (i.e., the ability of independent parties to test the system to ensure it doesn't unintentionally discriminate based on characteristics like race, religion, etc.) without complete transparency (i.e. publication of the algorithm) might be sufficient to protect the public interest, particularly if individuals or organizations who wish to access the APIs to audit the systems must first sign agreements not to use any data they derive during the course of the audit for competitive purposes.

22) Additionally, when it comes to AI systems, transparency should not be limited to just the algorithm or the code running the system; the datasets used to train an AI system are just as critical in order to ensure transparency. This is because datasets can have a tremendous impact on the performance of the AI system, causing problems even if the algorithm itself is flawless and unbiased.[3] Further, knowing what datasets were used to train an AI can help independent auditors discover where an AI system might be functioning in a biased or unfair manner.

23) As an aside, we close by noting that the dangers of black-box systems apply just as much to non-AI systems as to AI systems. A sentencing algorithm or a credit score doesn't have to use convolutional neural networks or other deep learning techniques in order to have a discriminatory or otherwise unfair impact on people's lives.

10. What role should the Government take in the development and use of artificial intelligence in the United Kingdom? Should artificial intelligence be regulated? If so, how?

24) In general, the Electronic Frontier Foundation is skeptical about government regulation of technology. Legislation often has unintended consequences and can easily interfere with the process of innovation. Most aspects of artificial intelligence are far too speculative and immature to be appropriate subjects for regulatory action at present. However there are some well-documented and serious problems with currently deployed machine learning systems for prediction and classification in institutional decision making.[4] They are urgent enough that careful, judicious, and consciously experimental regulation may be warranted in some domains.

25) For example, the problems of racial, gender, and other forms of demographic bias in deployed machine learning systems are so severe that they constitute serious public policy problems. Regulation requiring processes to at least measure and report, if not specifically correct,[5] such biases should be considered in cases where the decisions, choices, or recommendations these systems make significantly impact people's lives.[6]

[3] For example, see *Google Photos labeled black people 'gorillas'*, USA Today, https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/ for an example where a dataset that underrepresented black people resulted in unintentional results, or *Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day,* The Verge, https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist, for an example where an otherwise reasonable social AI system learned to be antisocial based on its interaction with Twitter users.

[4] To our knowledge, the real and well-demonstrated practical problems currently exist in comparatively simple systems such as regression models, rather than complex neural networks that would more fairly deserve the label "artificial intelligence". However it would be prudent to craft any regulatory principles so that they could apply to either type of system.

[5] Several methods for correcting bias have been proposed in the literature. For an entry to this literature, see M. Hardt, E. Price and N. Srebro, Equality of Opportunity in Supervised Learning, *NIPS 2016*, http://papers.nips.cc/paper/6373-equality-of-opportunity-in-supervised-learning, and https://research.google.com/bigpicture/attacking-discrimination-in-ml/. Since there are several incompatible standards for what fairness and de-biasing might mean (see J. Kleinberg, S. Mullainathan and M. Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores, *ITCS 2017* https://arxiv.org/abs/1609.05807) it may be prudent to require organisations deploying machine learning

26) On the topic of transparency and explainability, the EU is conducting a significant regulatory experiment with the “right to explanation” contained in the GDPR. Providing good explanations of what machine learning systems are doing is an open research question; in cases where those systems are complex neural networks, we don’t yet know what the trade-offs between accurate prediction and accurate explanation of predictions will look like.

27) In some domains of application, there are fundamental reasons for optimism about those trade-offs and therefore about the GDPR’s rules. For instance, where a classification system is trained on a set of readily describable input variables, it should be possible to provide good statistical explanations even if the classifier is a very complex neural network.[7] In other domains, particularly when the inputs are complex data like images are video, it’s not clear if accurate simple explanations of predictions will always be available, or what trade-offs will have to be made to obtain them.

28) Given these technical uncertainties, the EU’s “right to explanation” should be viewed as a regulatory experiment, and its successes and failures should be continually evaluated. To the extent that Brexit gives the UK additional flexibility in adopting the GDPR or otherwise, we would ask the question, “how can the UK take a position which builds on the utility of the EU’s experiment?” That might mean adopting clearer and stronger incentives for explainability, if the GDPR rules appear to be bearing fruit in terms of high-quality explanatory technologies, or it might mean moving in the direction of different types of rules for explainability, if that technical research program appears unsuccessful. Incorporating the EU’s own reviews and reforms would also be important.

29) Finally, we caution the government from enacting any regulations that are narrowly tailored to specific AI techniques.The right course when it comes to regulating AI is to focus on each of the AI system’s impact and domain of application, as opposed to the underlying technical methods. Regulations that focus on the technology, instead of the impact, will likely fail to protect the public, and they may also threaten innovation. As we noted in our answer to Question 9, a sentencing algorithm or a credit score doesn’t have to use convolutional neural networks or other deep learning techniques in order to have a discriminatory or otherwise unfair impact on people’s lives. Innovation in the field of AI is proceeding rapidly--both in terms of the

for high-stakes decision making to select and justify one standard while measuring and reporting their rates of deviation from the others. We would caution however that “maximising the accuracy of predictions” is rarely an appropriate notion of fairness. We would caution also that these bias mitigation techniques do not address the problems of building models from inherently biased data sources, and good regulation would find ways to incentivise companies to be skeptical of their training data and find ways to improve on or work around its flaws.

[6] We have no specific recommendations about how to delineate between high and low-stakes applications of machine learning, but rules that apply when decisions have an expected monetary value above some threshold are one obvious way to achieve this.

[7] For instance, the statistical explanation for a low insurance premium could be “In cases like yours, our model found complex relationships between age, gender and driving habits that predicted risk of an accident. Generally speaking, it helped that you were younger, female, that you started driving at a younger age, and that you rarely drove in high-risk locations.”

capabilities of AI systems and the scope of problems they can solve. But the field of AI safety research is also growing--both in terms of ensuring that AI systems act as expected and also, as we mentioned earlier, in terms of ensuring that AI systems act fairly. We urge the Government to focus on the end result of the deployment of AI systems when determining what regulations are appropriate.