# Written Testimony of Jack Clark
# Policy Director
# OpenAI

## HEARING ON

## “Artificial Intelligence: Societal and Ethical Implications’”

## BEFORE THE

## House Committee on Science, Space, & Technology

## June 26th, 2019.

## 1: Introduction.

Chairwoman Johnson, Ranking Member Lucas, and committee members, thank you for the opportunity to testify about this critical subject. This hearing is informed by my work at OpenAI, an artificial intelligence research and development company seeking to build general-purpose AI systems to benefit all of humanity. It is also informed by my work as a member of the Steering Committee for the AI Index, a Stanford initiative to track, measure, and analyse the progress and impact of AI technology.

When thinking about the ethical and societal challenges of AI, we must remember AI is a product of the environment it is developed in, and it reflects the inherent biases of the people and institutions that built it. Therefore, when we think about how AI interacts with society, we should view it as a *social system* rather than a technological system, and this view should guide the sorts of policies we consider when thinking about how to govern it.

For the purposes of this hearing I will discuss a relatively narrow subset of AI: recent advances in machine learning, oriented around pattern recognition. Some of these techniques are relatively immature, but have recently become 'good enough' for various deployment use cases. Crucially, 'good enough' isn't the same as 'ideal', and 'good enough' systems exhibit a range of problems and negative externalities which should require careful thinking during deployment. And whenever a system is "good enough" we should ask "for who?".

For this testimony, I will:

- Briefly outline recent progress in the field of artificial intelligence.
- Outline some of the ways in which contemporary and in-development systems can fail.
- Discuss the tools we have today to deal with such failures.
- Outline how government, industry, and academia can collectively address concerns around the development and deployment of AI systems.

### 1.1: Why we're here: We've entered the era of "good enough" AI

There are two classes of systems which are predominantly deployed today[1] - systems that classify the world according to an objective defined by a human, and things that predict something about the world and take an action. (As this hearing is predominantly focused on systems being deployed today or likely to be deployed in the future, I am limiting my overview here to the bits of AI which are gaining the most commercial interest.)

For classification, we have recently figured out how to create AI systems that can crudely mimic the capabilities of a couple of human senses: specifically, vision and hearing. By this, I mean

[1] Note that this description avoids discussion of 'expert systems' and other AI approaches which have been developed in prior decades and which have been deployed in parts of society since the 1980s. The focus of this testimony is on machine learning systems and specifically ones that primarily use deep learning - that's because these systems have broad capabilities and are being broadly deployed.

that recent advances in the field of 'machine learning' have let us develop systems that can - given a large enough dataset and computational power - learn to map labels to information extracted from images and audio. For instance, systems that assign a label to an image, or a part of one, like labeling fruit as being safe or rotten in farming, or a social network platform correctly identifying an individual in a photo, or a surveillance system classifying the actions of people in public spaces like train stations to identify suspicious activities.

To give a sense of the underlying pace of progress for this capability, we can look at the results of the 'ImageNet' object recognition competition: in 2010 a computer could be shown an image and, about **72%** of the time, come up with a list of five labels for the image, of which one would be correct. By 2017, this accuracy had climbed to above **97%**[2] - and progress is continuing.[3] This is progress on a particular dataset, but it relates to larger technological advancements, which loosely correlate to better performance on other specific tasks, like analyzing security camera footage, or spotting animals in nature. Similarly, for the field of speech recognition - that is, accurately transcribing speech - performance on one major benchmark has increased from **84%** in 2011 to **95%** in 2017[4].

However, these capabilities can degrade when exposed to things they haven't been trained on, like people of demographics different to the underlying dataset, or even products popular in "poor" countries.[5]

Meanwhile, research in reinforcement learning[6] has driven advancements in systems that can learn to act autonomously in specific circumstances. These systems can display their own patterns of failure, but it should be noted they are predominantly being research today, rather than widely deployed. (You can track the evolution of the capabilities of research systems here by looking at the complexity of the environment the agent can achieve an objective within. So, what does that look like? In 2013 we could use these systems to learn to play old Atari games like *Breakout!* and *Space Invaders*, in 2016 we could use such systems to beat humans at complex board games like *Go*, and in 2018 we could use these systems to compete with humans in very complex, real-time strategy video games like *StarCraft II* and *Dota 2*.)

The progress in these domains is impressive and worthy of attention, because they roughly correlate to contemporary or future societal impacts of AI: these performance increases, and associated ones in other domains, have led many AI systems to go from 'barely usable' to 'good

---

[2] Some research indicates that this exceeds human performance at this task. For more, see Andrej Karpathy "What I learned from competing against a ConvNet on ImageNet" http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/
[3] AI Index 2018 report, page 47. For more, see: https://aiindex.org
[4] AI Index 2017 report, page 31. For more, see: https://aiindex.org/2017/
[5] Does Object Recognition Work For Everyone?, DeVries et al: https://arxiv.org/abs/1906.02659
[6] Reinforcement learning is where you have an AI agent and a simulator (for instance, a flight simulator); you give the AI an objective (e.g, fly the plane from here to Spain), and then you have an AI system try to achieve this goal. The AI system will fail a lot, and each time it fails you restart the simulator and it tries again - eventually, the system will learn how to fly the plane to achieve the objective.

enough' in terms of real world deployment[7]. Given that AI is also a social system and has recently attained 'good enough' performance, we should ask good enough *for whom?* and good enough *at what?*

As my other panellists for this testimony will show, these systems, when deployed, frequently exhibit biases, and these biases can manifest as *inequitable access to the benefits of AI*[8] or *false positive identification by AI systems*. They also exhibit problems related to the process surrounding the design and deployment of AI systems, and some longer-term issues with the learning algorithms used to implement some AI systems.

We can expect progress in AI from both a research and a deployment view to continue, because of:

- Massive increases in the numbers of students involved in academic AI programs across the world.
- New funding from a variety of governments[9] and industry
- Falling costs of both computers and data storage systems.
- Ongoing algorithmic improvements.
- Commercial pressures; now that AI is "good enough" it makes economic sense for a large number of actors to invest in its development.

**1.2: AI progress and economic incentives**

Last year, OpenAI carried out an analysis in which we reviewed research papers relating to AI that were published in the last few years and analyzed the total amount of computational resources dedicated to the development of such systems. Our analysis showed that this amount had increased by **300,000X** over the past six years. The systems which fit this trend spanned use cases from image recognition, to machine translation, to strategic game playing systems. This trend correlates both to the increasing capabilities of some of these systems, and the increasing economic expenditures of large AI research and development organizations. (To put 300,000X in perspective, Moore's Law - that is, the 70-year trend that computers tend to double in capability every 18 months, would generate a **12X** increase over this same period.)

Many recent breakthroughs in AI systems for purposes like image recognition, speech recognition, machine translation, game playing, are correlated with this increasing compute

---

[7] If we were to define a turning point in this domain it might be around 2017 - that's when Google# (a subsidiary of Alphabet Inc.) described itself as an 'AI first' company, and other large companies signalled larger commitments to AI.

[8] For instance, research has shown that commercially deployed image AI systems from companies such as Amazon and others have significantly higher error rates at classifying females with darker skin tones. See: Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products, by Inioluwa Deborah Raji and Joy Buolamwini. http://www.aies-conference.com/wp-content/uploads/2019/01/AIES-19_paper_223.pdf

[9] Including, I hope, additional funding from the US government.

usage trend. This correlates to increasing economic expenditures by the companies deploying or researching the systems. This number also implies significant spending by people on the underlying computational systems required to train these AI systems, so the long-term trend could be altered by larger economic or R&D forces.

This number implies two things:

- 1) AI may progress more rapidly than peoples' intuitions would suggest, as people are bad at modelling what 300,000X increases correlate to.
- 2) We can expect the technical weaknesses of AI systems to 'scale up' with the amount of computational power poured into them, unless we develop smarter algorithms and better systems of governance for the organizations that develop them. This means that the ways AI algorithms fail at small scale can potentially be amplified and cause more harm when these failures occur in larger systems.

## 2. When 'good enough' AI goes bad.

I think there are two broad but related classes of failure we should think about here: when an AI system fails as a consequence of the *process* humans use when developing the system, and when an AI system fails as a consequence of the *learning algorithm* it has been equipped with[10]. For the purposes of this hearing, I think that failures of process are currently more numerous and consequential for society, while failures of algorithms may be significant in the long-term but are not as commonly seen in the wild today.

### 2.1 Process failures

Process failures typically manifest as an AI system failing dramatically during deployment, usually as a consequence of it being surprised by something. Unlike humans, AI systems are terrible at adapting to surprising situations, so these failures are typically severe as they speak to an underlying deficiency in the system. The system is typically surprised by something because it hasn't been built in a way that fully appreciates the context of the environment it is being deployed in.

Here are some examples of ways in which either researched or deployed systems have failed:

- Google's 'Google Photos' application incorrectly classified a black male as a gorilla. This failure was likely a consequence of the company not gathering enough data to teach its

[10] For a fuller overview of the various ways AI systems can fail - including systems currently on the frontier of AI research - please refer to "Concrete Problems in AI Safety" by Amodei et al (2016) https://arxiv.org/abs/1606.06565 and "Building safe artificial intelligence: specification, robustness, and assurance" by DeepMind Safety Research (https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1)

systems to consistently characterise black males and gorillas accurately, and not having sufficient testing regimes to identify this issue prior to deployment.

- IBM's 'Watson' healthcare system would sometimes recommend "unsafe and incorrect" cancer treatments, according to a report by STAT News, with the flaws emanating from improper dataset selection and improper processes for collecting people's opinions about what effective treatments were[11].

### 2.2 Learning algorithm failures

A good way to think about artificial intelligence systems and failure is that when they go wrong, it is usually because they achieved the specification but not the spirit of the described rule; these actions can frequently seem inappropriate or unsafe to a human. Sometimes these problems relate to weaknesses in the algorithm used itself, and other times they relate to humans mis-specifying the objectives of the algorithm.

- **Brittleness:** Image recognition systems can fail as a consequence of imperceptible variations in the appearance of digital and real images - images that cause them to fail are known within machine learning as 'adversarial examples'[12]. They can also fail as a consequence of dealing with unanticipated things - in one memorable example, researchers showed that by superimposing an image of an elephant onto an otherwise normal image, they could reliably cause image recognition systems to fail to classify other parts of the image[13].

- **Mis-specified rewards:** When training an AI system to complete a boat race in a video game, OpenAI gave the system the objective of getting as many points as possible, after observing that points typically correlated to winning the race. Our boat found a bug in the game that meant it could get a high score by navigating to a lagoon in the center of the race and spinning itself around to repeatedly hit various high scoring items, while setting itself on fire[14].

- **Mis-specified rewards:** When training a simulated robot to move its arm to move a hockey puck from one point of a table to another, o=OpenAI's robot instead learned to move the entire table to move the puck, rather than sliding it deftly, as we had intended. This would be dangerous in a real-world setting, and even if you installed safety systems

[11] For more, please refer to IBM's Watson Supercomputer recommended "unsafe and incorrect" cancer treatments, internal documents show, by Casey Ross for Stat News (2018). https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/
[12] For more, see 'Explaining and Harnessing Adversarial Examples' by Goodfellow et al, (2014). https://arxiv.org/abs/1412.6572
[13] For more, see The Elephant in the Room by Rosenfeld et al, (2018). https://arxiv.org/abs/1808.03305
[14] For more information, please refer to: https://openai.com/blog/faulty-reward-functions/

on the robot the fact this happens indicates other unanticipated behaviors could occur during training of the AI system.

- **Unexpected exploits:** Other examples abound, and are regularly collected and analyzed by the AI community. For example: A four-legged evolved agent trained to carry a ball on its back discovered that it could drop a ball into a leg joint and then wiggle across the floor without the ball ever dropping; in another scenario an agent chose to kill itself at the end of the first level of a game so it could avoid losing in level 2 of the game, and so on[15].

### 2.3 Process + Learning Algorithm Failures

Many failures occur as a consequence of process failures as well as learning algorithm failures. I think these situations are where many of the hardest problems occur, because they typically require a combination of technical and social analysis to understand and respond to. Some examples of failures of these types could include:

- Recommendation engines: Today, many companies around the world are seeking to use machine learning to learn to recommend products or services to people. When these systems fail it's usually a consequence of the underlying learning algorithm achieving a mis-specified objective (for instance, optimizing for engagement when showing people videos, which can lead to people consuming more videos that they find engaging, which can sometimes correlate to extremist content[16]), as well as the organization not doing enough direct study of the end effects on its users.

## 3. What can academia, government, and industry do to address these issues?

Technological fixes alone will be insufficient to address potential impacts of these technologies - this work will require careful coordination between industry, academia, and government during the development and deployment of these systems. However, a list of work without the accompanying resources to carry it out is useless, so I feel it is prudent for the government to consider increasing its own ability to measure, analyze, benchmark, and forecast the

[15] For many more examples, please refer to: https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/

[16] YouTube recently announced plans to remove thousands of extremist videos located on the web video service, according to *The New York Times* (June, 2019). https://www.nytimes.com/2019/06/05/business/youtube-remove-extremist-videos.html

development and application of AI systems, and to increase the funding it assigns to AI development[17] so academia is better equipped to solve these issues.

### 3.1: Government interventions

I think government has a profoundly important role to play here, chiefly by funding initiatives to gather more information about the progress and impact of AI systems. I believe it can step into this role via modest investment in its own capabilities to measure, assess, and forecast aspects of AI progress and impact. We need the equivalent of a publicly funded weather forecasting service for the ways in which AI is evolving so that we can better orient ourselves with regard to contemporary opportunities and problems and better spot problems and solutions that are over the horizon.

Specifically, I think government should intervene in the following ways:

- **Measurement, assessment, and analysis of deployed systems:** It would be helpful for the government to continuously benchmark for-sale or deployed machine learning systems for societally harmful failures, such as bias. Today, numerous academic researchers have developed datasets that deployed systems can be tested against; and we should consider building a 'bias test suite'[18], which government - potentially via agencies such as NIST - can develop as a resource for industry and academia.

- **Transparency in government AI procurement:** Today, it's difficult to get a sense for what AI systems are deployed[19]. Government can make a difference here by increasing the transparency with which federal agencies procure and deploy AI systems. This would equip academia with more information to use to study the impact of such systems, and would help further our knowledge about what responsible development and deployment of these systems looks like.

- **Funding:** We should increase the funding we allocate to artificial intelligence research and development in academia, while also increasing the resources to government agencies that can help coordinate actions between industry, government, and academia. I think that some existing proposed legislation, such as The Artificial Intelligence Initiative Act, could be helpful here. This legislation proposes increased funding for NIST, which would help that agency conduct more measurement and assessment of AI systems,

[17] This should be net-new funding for scientific research, rather than funding that detracts from existing research initiatives.
[18] Such a suite could consist of multiple datasets which systems can be tested against to show equitable effectiveness across a diverse set of people and objects.
[19] I have spent over two years working with the Steering Committee for the AI Index to gather data relating to deployment, and we've found the data to be piecemeal and partial. That's because there are few incentives or mechanisms to get people to describe the systems they deploy, and frequently the main way to know a company or government agency is using an AI system is via a press release from the vendor announcing them as a customer, through media reporting about the product, or through leaks.

while also creating more tools for the federal government to coordinate among itself as it further develops its AI strategy.

### 3.2 Academia

Academia should carry out more targeted research to deal with problems of process failures, learning algorithm failures, and the union of the two[20]. This will require a combination of directed technical research as well as heavily interdisciplinary research.

The main interventions I think would be useful here are[21]:

- The development of 'playbooks' in partnership with industry and government that can help AI developers avoid process problems.[22]

- Additional funding for interdisciplinary research that brings together multiple academic disciplines to analyze the contexts within which AI algorithms are developed and how these contexts interact with the technical aspects of the systems to cause problems.

- Continued funding for research that seeks to better understand the safety aspects of AI systems and to create tools to more easily interrogate AI systems for traits such as bias, or incorrigible behaviors.

### 3.3 Industry interventions

Industry, government and academia must engage each other more frequently and comprehensively. While this is a relatively obvious point to make, it bears repeating: I do not think our current conversations are as useful as they could be, nor are they as effective as they could be. My perception of why this is is threefold:

- Government lacks the technical expertise to provide enough touchpoints to industry and academia. By technical expertise, I mean people and institutions tasked with tracking and analyzing technical progress while also gathering data on societal impacts and discussing these findings with industry and academia.

- Academia rarely directly rewards policy engagement by younger students and junior faculty; typically, many tenure-track positions evaluate people for somewhat narrowly scoped work and achievements, and relatively few institutions would heavily weight

[20] We can enable such research via additional funding for academia.
[21] Many of these interventions are currently being carried out by academia, but my observation is that the scale of the issues are sufficiently large we should scale-up funding and activity here significantly.
[22] For an example in another domain, check out the US Digital Service's 'Digital Services Playbook' https://playbook.cio.gov/

policy contributions. (For example, a machine learning professor is predominantly evaluated today on their technical contributions, and typically via participation in the academic publishing system of peer-reviewed papers.)

- Industry tends to be cautious in its interactions with governments, especially when it comes to discussing some of the difficult questions surfaced by AI technology. Such caution makes sense when the government is perceived to lack sufficient personnel to have a detailed discussion, as there are reasonable concerns about misinterpretation leading to adverse policy outcomes.[23]

## 4. Conclusion

AI is progressing rapidly and, at the same time, it's clear that AI deployment brings both societal and technical challenges. We need to decide as a society what values we apply when developing “good enough” AI and where those values derive from, and we should continue to conduct technical work to give us the tools to better align these systems with societal preferences.

As discussed, I think what we need to address these challenges are:

- More transparency into systems that are being deployed into critical areas of public life.
- Increased government investment to measure, assess, track, and forecast the progress and impacts of AI.
- Greater efforts to make this an interdisciplinary conversation, as the problems are themselves interdisciplinary.

[23] Part of why I am being so blunt here is that the organization I work for does not deploy commercial products into the world, so there is less reason to be cautious during these conversations as we don't have a business that could be impinged on by regulatory actions in response to this testimony.