Written Testimony of Jack Clark
Policy Director
OpenAI

HEARING ON

"The National Security Challenges of Artificial Intelligence, Manipulated Media, and 'Deep Fakes'"

BEFORE THE

House Permanent Select Committee on Intelligence

June 13th, 2019.

Chairman Schiff, Ranking Member, and Committee Members, thank you for the invitation to testify about the national security threats posed by AI and its intersection with fake content and, specifically, "deep fakes".

My name is Jack Clark. I'm the Policy Director for OpenAI, an artificial intelligence research company in San Francisco, California. OpenAI's goal is to ensure that Artificial General Intelligence benefits all of humanity. To achieve our mission, we need there to be a stable national and international policymaking environment with regard to increasingly advanced AI systems, and we think that if we're able to respond to the challenges of AI-driven fake media, we'll be better positioned to respond to the challenge of other, future AI capabilities as well. I am also a member of the Steering Committee for the AI Index, a Stanford-led initiative to measure, assess, and forecast the application and progress of increasingly powerful AI systems. Additionally, before working in AI I spent around eight years working as a professional investigative technology journalist, so on a personal level I have some experience with both the production of media, methods used by media to assess the veracity of sources, and how we could expect media to respond to a landscape filled with increasingly dubious information.

This testimony will seek to situate deep fakes (and broader cases of AI-generated contents and synthetic media) within the larger artificial intelligence landscape, discuss some of our experience with the sort of AI research which synthetic media can spring out of, outline our view of contemporary threats and mitigations, and describe a combination of technical and regulatory actions which we think would make for a safer and more stable synth-media-world.

But, first, an uncomfortable truth inherent to this testimony: the problems of synthetic media are going to get worse, there are relatively few technical remediations that work in isolation, and society will need to internalize the notion that digital media is by-default untrustworthy (potentially via large-scale education of the public), unless accompanied by some kind of indicator of authenticity or verifiability from a trustworthy source[1]..

Part 1:

1.1: What are deep fakes and how do they relate to synthetic media and the broader field of AI research?

Just as PhotoShop has been used for many years to create fake images that, to most people, are indistinguishable from reality, contemporary techniques drawing on advances in cloud computing, machine learning, and data processing, mean it is becoming cheaper and easier to

[1] We could verify trust via technological tools, or via vetting from fact-checking or media organizations, or some combination of the two.

create ‘fake media’. This media ranges from things like ‘deep fakes’, to manipulated audio systems that imitate someone’s voice, to generative systems that can create or morph or edit images, video, audio, and text for a variety of purposes.

Much of the worries we have about the future of media relate to the increasing ease with which we’ll be able to cheaply create ‘fake’ rich media and use this to mount public opinion campaigns which could accentuate societal divisions, or cause political destabilization[2]. This is as much a societal problem as it is a technological problem, and we’ll need interventions in a variety of areas to better make ourselves resilient to the issues posed by these things.

When it comes to the technology underlying deep fakes and other fake media phenomena, it’s important to note how basic this technology is, and how fundamental its capabilities are to numerous, beneficial uses of artificial intelligence.

Faking stuff mostly involves taking data from one distribution and transposing it to another (e.g., making someone’s voice sound like someone else’s voice; making text that seems to be written in a certain style or with a certain worldview; inserting or deleting a person from a video, splicing one set of objects in one stream of data into another, etc). This same kind of operation is also used to *create* things, and since this kind of transpose is a fairly basic operation, we can expect the scientific community to develop numerous techniques to make this easier and more effective as they’re inherently useful - and we can expect such techniques to be adapted by bad actors to fake content. This is as much a historical trend as it is a technical one - the technology to record and edit audio or to manipulate images followed similar development trajectories, where early systems were expensive and complicated to operate, and over time they became simpler, easier to use, and more effective, and thus more bad actors found ways to use the cheaper and more widely proliferated technologies.

A key problem inherent to deep fakes and synthetic media in general is that the same technology overlaps with general-purpose tools and techniques used for other parts of research: many of the techniques these systems use are the same techniques you’d use to do things like: analyzing healthcare data; building the audio synthesis component of a speech translation and verbalization system; designing things for civic purposes (e.g., bridges and dams, or custom-fit medical devices); and playing a significant role in science by giving researchers new tools to analyze relationships between different domains.

Additionally, deep fakes and synthetic media, like other AI technologies, do not require specialized hardware or equipment to develop; a desktop computer with a powerful graphics card is sufficient (and affordable to the solo developer). The technology to make some of the things this hearing is concerned about is broadly intertwined with technologies and systems that

[2] For instance, there’s compelling evidence that now people are aware of deep fakes, they have already become more suspicious of video outputted by governments during fraught political periods; see the relationship between government-generated footage and an attempted coup in Gabon: https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/

power the economy and scientific enterprise of nations, and is also broadly diffused and available. We cannot expect the solutions to the problems of fake media to be simple, or easy to implement without further funding and study.

1.2: How available are deepfake technologies and associated tools used in the production of synthetic media?

The technology to make synthetic media in a variety of domains is widely distributed as a consequence of three factors:

- The AI development community has a general bias towards the creation, distribution, and circulation of open source tools and techniques to accelerate scientific research and discovery.
- The AI development community is currently discussing different norms around responsible disclosure and norms in AI research, and is similarly at the beginning of research about notions of responsible release and disclosure of AI capabilities.
- As mentioned, many of the technologies and infrastructures[3] that can be used to synthesize media, can also be used to do a far larger range of helpful things, ranging from tools for scientific discovery, to new ways to create art. A general trend over the course of history has been the diffusion of general-purpose tools.

To get a sense of just how available these tools are, while putting together this testimony, we took a look online to get a sense of the availability of the technology. We found a variety of 'Colab' notebooks online which would let us train various 'deep fake'-creating systems in our web browser, and found a larger number of open source repositories containing tools to make deep fakes on Github. We also found several commercial solutions where we could pay money online to generate various faked forms of media - but we didn't explore these in detail given the availability of so many free options.

Part 2: OpenAI's experience with fake media

2.1 GPT-2

Earlier this year, OpenAI revealed our second Generative Pre-trained Transformer ('GPT-2') system[4], which is a large-scale language system which can also generate text in response to a prompt.

The main thing to know about GPT-2 is we developed it as part of our technical research agenda toward artificial general intelligence. Specifically, for this project we were eager to

[3] For instance, people can rent powerful computers over the internet from cloud computing providers such as Microsoft, Google, and Amazon. These computers can be used to train and develop AI systems.
[4] More details here, and the linked web page provides a link to a technical report describing our system. https://openai.com/blog/better-language-models/

understand how we could take an existing well-understood part of AI (language modeling), develop a simple system incorporating some recently developed basic AI components, then train this system against a large amount of publicly available data and study what capabilities emerged.

The resulting large language model, GPT-2, has a wide range of capabilities which it learned without prompting by us (we exposed it to a large dataset and trained it with the objective of being able to predict the next word in a sentence). Some of its capabilities included: the ability to translate - at a very crude level - between different languages; summarization of articles; basic question-and-answer capabilities, and the generation of novel text in response to human-written prompts (e.g., “OpenAI testified in Washington on the 13th of June about how to deal with issues relating to the intersection of advancing technology capabilities and the creation of fake or misleading media, and…”, followed by a computationally generated prompt.)

In pre-release tests, we explored the generative capability of our language model by writing a range of prompts. We found out that if we gave it a text prompt written in a particular style (for instance, a news style, or the language used by people in online discussion forums, etc) we could increase the chance of it generating a particular output[5]. For instance, newsy prompts would be more likely to yield an article that would appear to be written from a news source, whereas a fictional prompt could yield something that read like a children’s story. Because of the range of capabilities it displayed, we became increasingly concerned about the potential ways in which this technology could be used for malicious acts, whether for the production of (literal) ‘fake news’, or to potentially impersonate people who had produced a lot of text online, or simply to generate troll-grade propaganda for social networks.

While confronting this intuition, we also noted that it was unclear to us how to systematically generate evidence about the nature of this threat, and how to weight the potential threat of such a technology against the broad utility of releasing it to the scientific community. Despite being somewhat uncertain as to the threat, we decided to be cautious in our approach to invite a conversation among AI researchers about publication norms, and to generate information about the creation of such norms through our own actions.

For this reason, for GPT-2, we have developed two release strategies that differ from the usual release approach of the AI community: Staged Release, and Partnerships.

By Staged Release, we mean that we chose to release the ‘small’ version of GPT-2 in February, then - following continued evolution in the wider AI research field - released the ‘medium’

[5] The level of control we can exert over the outputs of the text is fairly light at this stage, as is our ability to reliably generate outputs that read coherently. However, this technology will - like other AI technologies - improve over time, as a consequence of broader research by the scientific community.

version of GPT-2 in May[6], along with our thinking behind the additional release. We are currently evaluating release of further models and will be publishing our thinking here in August.

We chose Staged Release because it lets us slowly introduce a technology into the world, while being able to better monitor its usage and diffusion to calibrate our own threat model and systems of analysis to use when thinking about the distribution of increasingly powerful AI technologies. This also lets us embody a value we think is useful for AI researchers to consider - namely, choosing to gradually release technologies over time.

We are also exploring Partnerships, which is where we privately and non-commercially partner with other companies, institutions, and academia research groups so that we can share our larger GPT-2 models with them, so they can conduct research into mitigations and threat models and technical interventions.

By combining Staged Release and Partnerships we've sought to be more thoughtful in our approach to the release of an AI technology with a broad range of uses (some of which could be described as potentially abusive), while building out prototype 'new norms' (like Staged Release and Partnerships), which we are seeking to evangelize within the broader AI research and AI policy communities. We believe this is an area that would benefit from targeted funding towards interdisciplinary research. For example, it may be useful to fund continuous studies of how scientific communities in other areas deal with issues related to the anticipation of mis-use or abuse of their technologies[7], and also fund interdisciplinary workshops that bring together these researchers along with those of the government and AI communities to develop a shared language around threat models and threat-anticipating infrastructure. .

Part 3: What can we do about fake media?[8]

It seems like the following things are true:

- It is going to become easier to create increasingly convincing fake media as AI technology advances.

- AI technology is a general-purpose, omni-use technology, so it is challenging to call for specific technical controls to mitigate against specific outputs (e.g., deep fakes), without

[6] Our GPT-2 models come in various sizes, which correspond to the number of parameters in the model. Larger-scale models tend to be better at a variety of tasks and, in the case of text generation, as you increase the size of your models you tend to see increases in both the length and coherence of generated text, as well as the reliability of a given prompt yielding a good generation.
[7] Such communities could include those in: Gene-editing (e.g., 'CRISPR'), nuclear weapons and materials, hypersonics, gain-of-function research in biology, and others.
[8] Some of these recommendations overlap or have commonality with those recommended by Witness, an organization which supports people worldwide to use video and technology to protect and defend human rights. For more, see *Deepfakes and Synthetic Media: What should we fear? What can we do?,* published here: https://blog.witness.org/2018/07/deepfakes/

stifling broader innovation and research.

- Controlling the diffusion of technical AI capabilities is difficult-by-default due to incentives baked into the AI development community (and broader software development community).

- Many of the sources of control are not so much technical, as opposed to the systems that surround the technology. (For instance, one way to defend against people using AI-synthesized voices in telemarketing scams is to simply make it harder for criminals to spoof phone numbers).

Given these traits, we believe we need three types of interventions: technical, institutional, and political.

By technical, we mean there are a variety of specific technical interventions which can be made to help us *detect use of* these technologies.

By institutional, we mean there are things that can be done at the level of major technology platforms which could help to provide society with the ability to respond to large-scale fake media events.

By political, we mean that some interventions will occur at the level of government(s) taking actions, and these actions should likely include a mixture of building capacity in federal government, increasing dialogue between government actors and technical actors, and

3.1: Technical interventions

For technical interventions, there are a few avenues worth exploring:

Generator versus Discriminator research: The dynamic that plays out in a lot of these 'fake media' circumstances is a race between systems that can generate the fake and systems that correctly label the fake (or, in technical parlance, 'discriminate' it). And in the case of some specific AI technologies, improvements in one (e.g., advances in discriminators) can have a knock-on effect of improving generators as well.

It's unclear what the long-term dynamics of this are - many AI technologies work on the basis of optimization, and the sorts of technologies used to generate 'fake' media are optimizing for the objective of making something that tricks another AI system into believing it is real - therefore, since we can expect systems to trend towards being indistinguishable from reality, we should try to work out if it's possible for a discriminator to have an edge in the long run.

Combining detection systems with other signals: Though it may be possible to develop technologies that trick one discriminator, it's going to be more challenging to create things that can trick a discriminator as well as a system which, for example, studies the metadata

associated with the uploader to check for bot activity, along with other systems studying aspects of the user/uploader. Generally, taking a portfolio approach seems to be sensible, and we can imagine companies building and developing suites of AI-based and non-AI-based detection systems.

It is unlikely that any single technical solution will work as a permanent solution: as technology evolves, so too will the types of technical interventions needed to secure it. Therefore, it's valuable to create a consistent, long-term stream of funding for research into technical interventions here, potentially via expansion of existing initiatives (eg, work by IARPA or DARPA), or via net-new funding via broad disbursement mechanisms such as the National Science Foundation.

3.2: Institutional interventions

For institutional interventions, I think there are a couple of avenues of control:

Verified Human: Large social media platforms (e.g., Facebook, Twitter, YouTube), may want to provide direct authentication of content on their platforms. This can be done at a couple of levels:

- If the information was uploaded/posted by a person from the sensors on their smartphone, provide verification that the content originated on a device and was not manipulated before being posted to the social network.
- If the information was uploaded/posted by a person and it is unclear where it came from, provide a public, highly visible label that always says some variant of 'information provenance unknown'. If the platforms have implemented underlying systems to detect fake media, then may be worth additionally writing 'Our systems indicate this content is FAKE'[9].

Verified Image: For images and video which are edited, it would be helpful to create systems to log the lifecycle of these items - if I watch a video on the internet, it would be nice to know the provenance of that video, and whether it had been subsequently been edited by a third party who wasn't the creator. We could imagine baking in authentication systems to image and video editing software that could export a log of changes made to the image, giving us confidence that it hadn't been subject to undesirable manipulations. Work here could be linked to broader work by social platforms, letting us identify the lifecycle of a given piece of content, and better defend against outside actors manipulating it for malicious purposes.

Large Models with Staged Release: It may make sense for social media platforms and news organizations to develop the capacity to privately train and share large discriminators and

[9] Such a message could provide valuable information to malicious actors developing systems designed to frustrate such fake-detectors, so I would defer to experts in information security as to what should be disclosed here.

generators with each other so they can collectively understand the altering capabilities of synthetic media technology, and be better able to anticipate and mitigate against its uses.

3.2: Political interventions

We need political interventions here, because as this testimony has hopefully shown, the problem of fake news requires technological and institutional responses. Fundamentally we're talking about truth and how we approach truth in our peculiar modern era. For questions like this, I want the government to be involved, because making choices about truth and how we value it and protect it is connected to the overall health of a society - therefore, elected officials should be involved in these conversations. (Note that 'truth' and 'speech' are different here, so I'm focused on interventions to increase our ability to have information that is deemed true (as opposed to interventions to increase the likelihood of *speech* being 'true', which seems to be a path laden with risks and ambiguity).

My suspicion is that as the technology evolves we'll want to consider large-scale regulation to deal with some of the likely issues, but I think today we need to concentrate on actions that give all media consumers better information about the entire intersection of fake media, artificial intelligence, and digital platforms. My belief is that one of the key challenges of AI from a policy perspective is the rate with which it progresses in capability, which typically outruns the response pace of policy infrastructures: we should change this, and I think a prerequisite to rapid response is knowing about the thing you're responding to.

I think there are several interventions which can be done at the level of federal government. These include:

Measurement and Benchmarking systems for synthetic media: How good are various systems at generating synthetic media outputs, and how well can existing technical systems recognize truth from reality? The short answer to this question is: we don't know! The long answer is: the AI community has developed a variety of different techniques that can be used to algorithmically judge the quality of the outputs of generative models, but these techniques are not standardized and are unlikely to become so in the future.

Government has a valuable role to play here in convening people from across academia, industry, and government, to discuss shared measurement techniques which can be used to assess the ongoing advance in capabilities of these systems, and to support the development of standards for assessing this advancement. The creation and maintenance of such metrics and standards would give government the ability to orient its policy responses according to the technology's contemporary performance as well as its likely evolution. A dedicated initiative to build a bench of talent inside government focused on measuring and assessing AI progress in domains deemed important to national security is crucial to closing the response gap

Comprehensive AI education: It would be helpful for the government to invest in a variety of education schemes so that people (kids and adults) can easily familiarize themselves with AI and develop intuitions about this technology[10].

Fake media research challenges: Government should continue to invest in media identification programs like DARPA's 'Media Forensics' initiative, and may want to increase funding here. Companies should seek to share their strategies and self-developed tools for combating such fakes, potentially via private disclosure to people from academia and policymakers; we could also incentivize the public sharing of tools via government-backed competitions aimed at developing tools to counter, track, and analyze fake media.

3.3 Further suggestions from a survey

When preparing for this testimony, I requested feedback from the broader AI community for input. At the time of filing, I had received 25 responses from people from industry and academia and other professions, including journalism. Some of the common threads in the responses were:

- Fake media and deep fakes are as much a media problem as they are a technological problem.
- It is not productive to regulate basic AI research, due to its manifold beneficial uses.
- Context matters: sometimes the same AI tool can be used for good purposes as well as bad ones. We need to focus on understanding and minimizing incentives for bad actors and developing tools for good actors.
- We need to invest in tools to assess the state of fake media and to identify it in-the-wild.
- We should invest in shared systems that can authenticate the validity and provenance of data, like images and videos and audio recordings.
- We should explore technical approaches to 'watermarking' systems for tools used in the generation of fake media.

4. Conclusion

As I hope this testimony has shown, society must prepare for a world where it is cheap and easy for a large number of people to perturb or fake media that is becoming increasingly difficult to distinguish from the truth. We should make significant investments in initiatives to track and assess the development of technologies that shape this landscape, understand the potential threats both to the US, and to broader international stability, as well as building tools to help us facilitate tracking developments here. This will require numerous discussions between academia, government, and the private sector. In addition, we should continue to invest in

[10] For instance, as part of the country's Finnish Center for Artificial Intelligence initiative (https://fcai.fi/), universities and companies worked together to create the 'Elements of AI' (https://www.elementsofai.com/faq/who-created-this-course), a free online course designed for non-technical people to take to familiarize themselves with the technology.

understanding the norms and practices that shape the AI research community, and think about how such norms could potentially be changed in the future to increase overall societal safety without sacrificing scientific progress.