Written Testimony of Jack Clark
Co-founder, Anthropic. Co-chair, AI Index. Member, National AI Advisory Committee.

Before the
U.S. Senate Committee on Commerce, Science, and Transportation

Thursday September 29th, 2022
Securing U.S. Leadership in Emerging Compute Technologies.

How testing and experimental infrastructure will let the United States take advantage of the industrialization of AI.

Chair Cantwell, Ranking Member Wicker, and members of the committee, thank you for the opportunity to speak with you today about the important topic of how the United States can maintain its leadership in emerging compute technologies. First, thank you for passing the CHIPs and Science Act. Through passing this, you have helped to set the United States of America up for continued leadership in the development and deployment of transformative technologies such as artificial intelligence.

For this testimony, I will make a few simple recommendations, which I hope will help us meet the challenge of international competition; increase opportunities for collaboration between the government, industry, and business sectors; and build a diverse and inclusive workforce to meet the growing demands of our evolving economy – all while ensuring the safe and responsible adoption of technology.

These recommendations are as follows:

- The United States should fully fund the CHIPS and Science Act, and make further investments in the measurement and monitoring of the artificial intelligence development ecosystem both domestically and abroad. Having accurate information about progress within the United States, among our allies, as well as progress occurring in other countries, is crucial for making good decisions about American technology strategy. We need to know if we're in the lead or if we're coming from behind, and where any gaps may be.
- The United States should seek to develop experimental infrastructure at scale for the development and testing of artificial intelligence systems by academic and government users. Concretely, the proposed National AI Research Resource can be best leveraged by pairing it with the creation of testbeds which can be used to train a new, diverse workforce in the important work of developing and assessing AI systems for economic applications and safety assurance.
- The United States should prioritize the development of resources for the assurance of AI - specifically, tools for the testing, evaluation, and benchmarking of artificial intelligence systems. The better we get at AI assurance, the more confidence we can have in AI systems, and the more we can create opportunities for collaboration across the private

sector, government, and academia. Additionally, as more communities have the ability to test out different applications, they'll develop new products and services along the way.

Before I expand on these recommendations, I'd like to state why they're necessary, and why I'm so appreciative you are having a hearing about this now.

HOW WE GOT HERE

First, I want to provide an update on just how rapidly the field has been advancing. The past decade has been distinguished by what we can think of as the industrialization of artificial intelligence; AI has gone from an interesting topic of research and discussion among researchers, to something of real economic and strategic utility.

We can very roughly draw the "ignition point" for the industrialization of AI to 2012: this is when a team of researchers at the University of Toronto were able to win a highly-competitive image recognition competition known as ImageNet using a then-novel technique — taking a bunch of so-called neural networks, layering them on top of one another like a lasagne, and then training them on a significant amount of data[1]. The result was a system which set a new state-of-the-art on image recognition and which led to significant investments in AI by industrial actors here and abroad.

Since then, the field has notched up a few notable achievements. Systems like "AlphaFold"[2] have revolutionized the field of protein structure prediction, which is a key input to scientific development. Other AI systems have proven better able to stabilize the plasma in fusion reactors than any human or previous software system[3]. Meanwhile, AI has begun to make its way to the consumer so quickly that many do not realize they're already interacting with it throughout the day: voice recognition systems have improved substantially, we're all able to search through the photos on our phones now to find pictures of our dogs, cats, and family members.

These developments are fantastically exciting. Remember, ten years ago, none of these things were possible. Now they are. AI is now being applied in a vast range of fields, and we've barely scratched the service. Just to give you an idea of what is possible, here are a few examples of how AI is being applied today and the positive impact it is having on the world:

- **Enhancing developer productivity:** GitHub's Copilot product – an auto-completion tool used for computer programming tasks – has been shown to make software developers

[1] ImageNet Classification with Deep Convolutional Neural Networks, https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html, 2012. One of the team members, Ilya Sutskever, went on to help found OpenAI, a major AI research company based in the United States.
[2] AlphaFold: a solution to a 50-year-old grand challenge in biology, https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology, 2020.
[3] Accelerating fusion science through learned plasma control, https://www.deepmind.com/blog/accelerating-fusion-science-through-learned-plasma-control, 2022.

more than 50% faster in their work than developers that do not use the tool. The study also found higher levels of developer satisfaction, with 60-75% of users feeling less frustrated by daily programming tasks[4] (GitHub).

- **Estimating sustainable development outcomes:** Researchers at Stanford University have demonstrated how combining satellite imagery and machine learning can help measure sustainable development outcomes in areas such as hunger relief, population density, and economic activity[5] (Stanford, Science)
- **Measuring agricultural health**: Academic researchers have developed an image recognition system for detecting agricultural diseases in the cassava plant (a critical food source for millions of people across Africa), using just a mobile device[6] (arXiv).
- **Low-resource language translation:** Google researchers have found a way to develop translation technology for underrepresented languages using only text in the original language (traditional machine translation systems typically work with two sets of text: the original language text, and its translation to the target language). Using this new approach, Google was able to add 24 under-resourced languages to its Translate service and develop a repeatable method to include other languages from around the globe[7] (Google Research, arXiv).

There are also exciting developments at the frontier; in the past few years, so-called "Foundation Models"[8] have emerged which show how AI is moving from an era of dedicated and specific tools to models that behave more like "Swiss Army knives" - a single model will be able to do a broad range of tasks, many of which are scientifically and economically useful.

These models are distinguished by the sizes of their datasets (extremely large datasets, ranging from hundreds of thousands of audio samples[9], to hundreds of millions of images[10], to billions of text documents[11]), the amount of computation required to train them (hundreds to thousands of specialized AI-training computer chips, running for months), to the complexity of the neural networks (which now number in the tens to hundreds of billions of parameters). Foundation Models have already proven to be useful: they can write and compose code, edit text, produce

---

[4]Quantifying GitHub Copilot's impact on developer productivity and happiness, https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/, 2022.
[5] Using satellite imagery to understand and promote sustainable development, https://www.science.org/doi/10.1126/science.abe8628?cookieSet=1, 2021.
[6] See: Using Transfer Learning for Image-Based Cassava Disease Detection, https://arxiv.org/abs/1707.03717, 2017.
[7] See: Building Machine Translation Systems for the Next Thousand Languages, https://arxiv.org/abs/2205.03983, 2022.
[8] On the Opportunities and Risks of Foundation Models, https://fsi.stanford.edu/publication/opportunities-and-risks-foundation-models, 2021.
[9] Introducing Whisper, https://openai.com/blog/whisper/, 2022.
[10] Revisiting Unreasonable Effectiveness of Data in Deep Learning Era, https://arxiv.org/abs/1707.02968, 2017.
[11] An empirical analysis of compute-optimal large language model training, https://www.deepmind.com/publications/an-empirical-analysis-of-compute-optimal-large-language-model-training, 2022.

images, edit images, summarize documents, form the basis of question-and-answer systems, serve as potentially useful educational tools, and more.

However, the frontier is expensive: it costs millions to tens of millions of dollars to train these models and therefore they are being developed by only a small set of predominantly private sector actors. A challenge we must overcome is how to broaden the experimental infrastructure necessary to investigate these models, so that more Americans can participate in the development and analysis of them, and also learn the engineering and research skills they require.

These examples illustrate how broad the effects of the industrialization of AI are. But if we want to capture all the upside of this technology and mitigate its downsides, we also need to think about policies and investments that can support the burgeoning ecosystem, and assure the safety and reliability of the systems being developed within it.

WHY TESTBEDS, DATASETS, AND EVALUATION UNLOCK AI INNOVATION

While there are many reasons to be optimistic about the potential opportunities afforded by AI, there are also well-documented risks[12] and biases[13] inherent in many of today's applications. These risks and biases make it harder to deploy safe AI systems, and because these risks and biases are hard to identify, they can also lead to AI systems being deployed which have inequitable or harmful behaviors. However, we can mitigate these issues through pre-deployment and post-deployment testing, both of which the government can support - specifically, via the National Institute of Standards and Technology (NIST).

To maximize the potential of AI technologies, one important role the government can play is in developing a robust ecosystem for AI assurance. An assurance ecosystem allows multiple stakeholders to assess AI systems for performance and safety through a combination of shared testbeds, datasets, and evaluations. System assurance provides model developers with certainty in the reliability of their models, end users with trust that models will act as intended, and government stakeholders with confidence that systems are safe for the general public. We can imagine this assurance ecosystem as being analogous to how product safety standards give consumers confidence in things ranging from cars, to food, to drugs. It's definitely time to build out this ecosystem for AI.

Beyond improving the safety and reliability of AI systems, shared testbeds and evaluations enable a stronger R&D environment. Generally speaking, whenever a set of researchers create an artificial intelligence model, they then run that model through a large-scale battery of tests to assess model performance against previous iterations, as well as other results in the public

[12] The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, https://arxiv.org/abs/1802.07228, 2018.
[13] What Do We Do About the Biases in AI?, https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai, 2019.

domain. External benchmarks provide an objective, baseline measure from which developers can compare and improve their systems.

At the same time, sometimes models are found to have capabilities that their developers did not anticipate, typically through end-users running novel or unexpected tests.[14] These tests can reveal both new capabilities as well as safety issues - for example, when GPT-3 was released, external users discovered that the system was able to perform some basic computer programming tasks as well as text-based tasks. Similarly, a few years ago, external researchers discovered that commercially deployed facial recognition systems displayed harmful biases, via a study named “Gender Shades”.[15]

We already have examples of how these kinds of testing methodologies can be operationalized; following the publication of Gender Shades, NIST significantly updated its Facial Recognition Vendor Test (FRVT)[16] to include fine-grained, granular evaluations which were also sensitive to the demographic makeup of potential end-users of the system. This highlights how you can operationalize testing in a way that improves both the safety of the system (by reducing the likelihood of deploying unfair systems), and also giving confidence to end-users of the system that it is going to perform well.

Given the passage of the CHIPs Act and the funding it seeks to allocate to NIST, we should consider all the ways NIST can play an expanded role here. What might it look like to identify areas where industry and academia would benefit from more robust tests and to seek to create them? How might NIST construct fact-finding teams to identify some of the areas of greatest “evaluation need” and create tests in response? And can we take the in-development NIST AI Risk Management Framework (RMF)[17] and identify specific evaluation methodologies or tests that we might invest further in, so as to unlock even more economic innovation and increase the safety of such systems? (In my day job at the AI research company I am a co-founder of, Anthropic, I spend a lot of time trying to better evaluate our systems, and I can tell you that we generally try to incorporate all the tests that exist outside of Anthropic into our testing framework. We really can’t get enough of them.).

These examples highlight the value of testing for both economic expansion, as well as improving the safety and reliability of AI systems.

WHERE OTHER NATIONS ARE

AI research and development is global. Most data sources tell us that, outside America, other key countries for AI R&D include the United Kingdom and, most pertinently for the field of international competition, China.

---

[14] Predictability and Surprise in Large Generative Models, https://arxiv.org/abs/2202.07785, pg 4, 2022
[15] Gender Shades, https://www.media.mit.edu/projects/gender-shades/overview/, 2022
[16] Ongoing Face Recognition Vendor Test (FRVT), 36th edition of the report, https://pages.nist.gov/frvt/reports/11/frvt_11_report.pdf, 2022.
[17] NIST, AI RISK MANAGEMENT FRAMEWORK, https://www.nist.gov/itl/ai-risk-management-framework, 2022.

China and the US can, in many ways, now be considered at roughly the same point in AI development. Both countries approach the development of the technology with different strengths and weaknesses that stem from their differing political structures, but both host burgeoning ecosystems of commercial AI companies, and both are supported by strong academic research infrastructure. The data bears this out:

- For example, while the US leads in the number of global AI research paper conference citations each year (30%, over 15% from China in 2021), China continues to lead the world in the total number of AI *publications* (journal, conference, repository combined), with over 60% more than the United States in 2021 (2022 AI Index)
- As it relates to patents on AI technologies, China now files over half of the world's patents (51% in 2021). The US still leads the percentage of *granted* AI patents globally, but that percentage has, on average, decreased over the past 7 years while the percentage of granted patents from China has steadily increased (2022 AI Index).

Beyond the basic metrics of academic publication, there are some qualitative examples I can share that illustrate how China has begun to advance its research and development of artificial intelligence.

ImageNet: Chinese teams became increasingly competitive at the aforementioned "ImageNet competition" in the 2010s. One Chinese team even won an image recognition challenge within that competition several years ago[18]. This is an extraordinarily competitive competition and being in the top-3 placed teams was typically considered impressive, and winning it is a proxy signal for competence. Put plainly: you win ImageNet by being extraordinarily good at training image recognition systems.

GPT-3: In 2020, an American AI research company called OpenAI (I worked there at the time) published a paper on a system called GPT-3. This system was a so-called large language model (LLM). LLMs are interesting to AI researchers because they are generic AI systems, capable of classifying and generating arbitrary text, and performing a broad range of tasks. LLMs are also distinguished by their cost: GPT-3 cost a lot of money to train, and involved using a large number of training accelerators (in this case, graphical processing units) to train a single neural network model; it could be considered a frontier capability due to this expense and complexity.

After publishing the paper about GPT-3 in May 2020[19], the first public replication of the system arrived in a paper in April 2021. The replication was a system called Pan-Gu and was developed by the Chinese telecommunications company Huawei[20]. Other replications followed

[18] Large Scale Visual Recognition Challenge 2016 (ILSVRC2016), https://www.image-net.org/challenges/LSVRC/2016/results.php, 2016.
[19] Language Models are Few-Shot Learners, https://arxiv.org/abs/2005.14165, May 2020.
[20] PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation, https://arxiv.org/abs/2104.12369, April 2021.

(HyperCLOVA from Naver in Korea, and after that Jurassic-1-Jumbo from AI21 Labs in Israel). Notably, two more replications from Chinese labs followed in Huawei's footsteps, and this year a research group linked to Tsinghua University released GLM-130B, a GPT-3-style language model which is currently the best-performing language model[21] available as open source - and it's made in China.

Re-identification surveillance: China is far ahead of the United States in the development of surveillance technology. Specifically, we can look at re-identification; the task of identifying a person in security camera footage, then being able to see that person via a different security camera posed at a different angle and use a machine learning system to figure out it is the same person. This is a powerful and chilling capability which violates the norms and privacy protections we have in the United States. However, just because we wouldn't necessarily adopt a technology ourselves, it's worth noting when someone else is ahead on a given capability, no matter how distasteful. In re-identification, recent research shows that Chinese teams have continually pushed forward the state of the art, and are responsible for 58% of the top papers in the field[22]. Re-identification requires a combination of large datasets, the development of large-scale neural networks, and creative algorithm design. In other words, being good at re-identification means that you've built a decent AI competency, and it's worth noting that China is ahead here.

HOW THE UNITED STATES CAN SOLIDIFY AI LEADERSHIP

Besides continuing to invest aggressively in fundamental research, the United States has a couple of strategic policy investments it can make to bolster its leadership in artificial intelligence. I've already discussed the importance of making investments in testbeds, datasets, and evaluation to further unleash US innovation here.

An additional lever we can use is the provisioning of experimental infrastructure for our academic community. Specifically, we should make it easier for America's best academic researchers to access computational power close to that found in industry, so that our universities can carry out ambitious experiments near the frontier of AI research; while it's likely industry will continue to define the frontier due to the increasingly large-scale resources being invested in model training, we should ensure academia is able to conduct experiments sufficiently close to it that they can help with the important work of safety validation and analysis of cutting-edge systems. This provides both accountability for the private sector, as well as letting our academic researchers design tools and techniques to improve systems deployed in the economy. We also need the necessary infrastructure to build testbeds for these systems so more people can spend time working out how to unlock their economic opportunities, and we can use these testbeds to include a much broader group in the testing and development of AI, which should help us grow the future AI workforce.

---

[21] For a detailed breakdown of performance, please refer to this GitHub page for the model: https://github.com/THUDM/GLM-130B, 2022.
[22] Measuring AI Development A Prototype Methodology to Inform Policy, https://cset.georgetown.edu/wp-content/uploads/Measuring-AI-Development.pdf, 2021.

For this reason, Anthropic firmly supports the goals of the National AI Research Resource ("NAIRR")[23] — a shared, public research infrastructure for academic researchers. We view its establishment as a necessary and excellent long-term investment in American AI research, as well as a critical resource for supporting the training and testing of AI systems. Increasing academic access to the infrastructure necessary to train increasingly resource-intensive models will build on the long and successful collaboration between academia and the US government in creating transformative technologies and advancements across the US economy. (We should also note that other parts of science already build large-scale experimental infrastructure, such as the particle physics community.)

From the 1960s until 2010, research shows that academia represented the majority of large-scale AI experiments. Between the early 2010s and today, the level of compute required for the largest scale experiments has increased by more than 300,000x — and the industry-academic balance has altered, with the vast majority of large-scale results now being carried out by industry rather than academia[24]. Few academics can afford the computing and engineering costs required to build and study large-scale AI models, such as Foundation Models which can cost millions of dollars to develop, and this is preventing some of our best researchers from working on problems found at the frontier. Even in Canada, where the country's advanced research computing (ARC) platform has allocated increasing quantities of compute to academics since the mid-2010s, the number of new applications for compute by Canadian researchers has grown at >10% a year – and demand still outstrips available supply.[25] Given academia's role in evaluating the safety and societal impacts of new technologies, we expect resources provided by the NAIRR will help restore a healthy balance between American universities and companies in cutting-edge AI research. Testbeds and other programs funded by the NAIRR will enable AI safety research and other research agendas in the public interest.

It's reasonable to ask why something like a NAIRR is necessary, given that AI is being developed and deployed by a large range of industry actors. After all, we might ask, isn't this a sign that the government should concentrate its efforts elsewhere? The answer is that by developing and funding a NAIRR, we're able to build infrastructure that will naturally serve as a proving ground for some of the ideas coming out of academia (and perhaps not yet mature enough to be adopted by industry), as well as creating infrastructure which is highly complementary with the testbeds NIST is tasked with building as part of the CHIPS and Science Act. Concretely, we might imagine the NAIRR serving as a resource for universities to develop large-scale AI systems, then we can also use the computational power of the NAIRR to facilitate a broad spread of universities to run testbeds to see how well we can turn these systems to often neglected problems; improving the way we manage our farms, building sensing systems to help us respond to natural disasters, figuring out ways to make our transport and logistical

[23] THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH RESOURCE TASK FORCE (NAIRRTF), https://www.ai.gov/nairrtf/.
[24] Addendum: Compute used in older headline results, https://openai.com/blog/ai-and-compute/, 2019.
[25] , 2022 Resource Allocations Competition Results, Digital Resource Alliance of Canada, https://alliancecan.ca/en/services/advanced-research-computing/accessing-resources/resource-allocation-competitions/2022-resource-allocations-competition-results

systems more efficient, and so on. This would also be directly enabling for another key aspect of CHIPS and Science – the NSF's new Technology, Innovation, and Partnerships directorate.

We suspect many of the best ideas for how to harness AI are going to come from universities across America working to solve problems relevant to their own communities, giving more institutions a role in assuring that AI systems are safe, reliable, and well-calibrated to the problems they are designed to solve. Local context is vital to ensure tools are well-designed. AI logistics systems will best serve the Port of Gulfport or the Port of Seattle if local port employees and nearby universities help inform those systems. These tools will underperform if they are only pressure-tested in distant labs, rather than by researchers close to local contexts.

Another example is in healthcare: AI shows significant promise to enhance healthcare, such as applications that aid in diagnostics or recommending treatments. However, due to differences between hospitals, such as imaging equipment, procedures, and local population demographics, AI algorithms that may work well at one hospital can perform very poorly at another. One solution to this problem involves more collaborations between universities and their local hospitals to develop, test, and fine tune models to serve regional needs, which can be facilitated by testbeds in combination with the NAIRR.

The NAIRR can also help us work on some of the challenges of the governance of increasingly capable AI systems. By making available experimental infrastructure for large-scale experimentation, the NAIRR creates an opportunity to think about how we govern that experimental infrastructure. Which experiments should get authorized for using a large amount of NAIRR resources?[26] How do we test and evaluate the systems that result from these experiments?[27] Which people should participate in the analysis and curation of the datasets which are used to develop models on the NAIRR? Once a system is developed on the NAIRR, how might organizations such as NIST help assure the resulting system for safety? These are all extraordinarily valuable things to work on in public, rather than in private as is done today by most industry actors. Beyond enhancing economic competitiveness and safety, the NAIRR may also help us identify smart, lightweight regulations that will be fit for the increasingly powerful AI models that will distinguish this new period of industrialization.

CONCLUSION

In conclusion, I'd like to thank this committee for its important role in the CHIPS and Science Act, enthusiastically support full appropriation for its programs, and recommend that the additional investments that the bill proposes be made in testbeds, datasets, and evaluation as a means of unlocking economic innovation and unlocking revolutionary scientific research.

---

[26] Centre for the Governance of AI Submission to the Request for Information (RFI) on Implementing Initial Findings and Recommendations of the NAIRR Task Force, https://www.governance.ai/research-paper/submission-nairr-task-force, 2022.

[27] Anthropic response to Request for Information (RFI) on Implementing the Initial Findings and Recommendations of the National Artificial Intelligence Research Resource Task Force https://www.ai.gov/rfi/2022/87-FR-31914/Anthropic-NAIRR-RFI-Response-2022.pdf

Additionally, I want to offer strong support for the establishment of the National AI Research Resource, a national infrastructure built to facilitate academic AI research that can also help make sure the United States stays ahead in one of the most important technology developments we are seeing today. The NAIRR will ensure that American scientists keep our nation at the forefront of understanding frontier compute technology, creating important research and workforce opportunities and new avenues for economic growth.

Thank you for the chance to speak. I'll be happy to answer any questions you have about my testimony.[28]

[28] I used an Anthropic 'Foundation Model' to write the 'CONCLUSION' section of this testimony. I added some specific language and tweaked a couple of words, but other than that, the text is what the AI generated. I hope this illustrates how these technologies are already changing how we work today.