BOTS OVER BRUSHES:

THE LOOMING COMPETITION BETWEEN GENERATIVE AI AND SMALL CONTENT CREATORS

Prepared by
the House Committee on
Small Business Democratic Staff

Committee on
SMALL BUSINESS
Ranking Member Nydia M. Velázquez

Table of Contents

Executive Summary 3

Background: Generative Artificial Intelligence (GenAI) 4

- Overview 4
- Underlying Technologies 5
  - Machine Learning 5
  - Neural Networks 5
  - Foundation Models 5
- Capabilities 5
  - GenAI Can Make Art 6
  - GenAI Can Creatively Write 8
  - GenAI Can Make Audio and Music 9
  - GenAI Cannot Think for Itself Yet 10
- Training Data 11
  - GenAI is Trained on Highly Expansive Datasets 11
  - GenAI Outputs Are Significantly Impacted by Items Included in Training Datasets 12
- “Black Box” Issue: GenAI Decision-Making and Training Processes Are Nontransparent 13

GenAI Has Serious Ramifications for Small Creators 13

- GenAI Competes with Human Creators and Devalues Their Work 13
- GenAI Uses Human Creations to Learn and Mimic, Unbalancing Competition with Human Creators 14
- “Black Box” Issue Obscures Potential Harms and Risks 15
- Ongoing Litigation and Need for Congressional Action 15

Policy Possibilities 15

- Mandated Disclosure of Unlicensed and Copyrighted Content in Training Datasets 16
- Licensing and Compensation for Use of Content in Training Data 16
  - Opt-Out Systems and Do-Not-Train Tags 17
  - Clarifying Fair Use Exceptions 17
- “Gold Standard” Training Datasets 17
- Watermarking GenAI Creations 18

Conclusion: GenAI Policy Must Benefit America’s Most Vulnerable Content Creators 19

Executive Summary

Generative artificial intelligence (GenAI) is a kind of AI software that can make written, visual, and audio content for users upon request. With GenAI software, users can specify what kind of content they are looking for in writing and receive outputs that are statistically likely to fulfill their requests. This software, which is based on the human brain, can do this because it can identify, understand, and replicate patterns among examples of the target medium. The process where a GenAI software scans these examples to determine what patterns exist is called training. When a GenAI software trains, it typically requires millions, or billions, of examples.

Assembling a dataset containing such a large quantity of content examples is highly resource intensive. Because of this, many GenAI firms do so by extracting content from as many publicly accessible websites as possible, a process called scraping. However, these firms often scrape the internet without licensing content from their owners.

The increasing popularity of GenAI, which can generate content faster and cheaper than America's small creators, poses a serious threat to their livelihoods. GenAI firms' unlicensed use of these creators' work compounds the intensity of this competition and adds layers of unfairness. Small creators are not only having their work used without compensation, but also seeing their work improve the software that could push them out of business.

This report seeks to address this issue by highlighting relevant policies being developed across the House of Representatives and examining their implications for American small creators. Four kinds of proposals have received considerable attention: (1) a statutory mandate for GenAI firms to disclose the unlicensed contents of their training datasets; (2) creating licensing and opt-out schemes for the use of online content in GenAI training; (3) government-made training datasets; and (4) watermarking GenAI outputs. Each of these four policies would bring benefits to America's small creators, but in different ways. First, the disclosure mandate would help small creators know if their work has been included in training data, allowing them to understand if they have been harmed and if they can seek legal recourse. Second, the licensing scheme would assist small creators in receiving compensation for the use of their work and empower them to prevent the use of their work in GenAI training. Third, the government training dataset proposal could also be helpful, though some controversial questions regarding censorship and misinformation would need to be resolved before the proposal is further developed. Lastly, the watermarking scheme would help content consumers better understand where they are receiving content from, and address aspects of the competition issue by helping distinguish GenAI and human creators' work in the broader marketplace.

As the legislative conversation widens and intensifies in Congress, it is important to keep the interests of America's small creators in mind. As GenAI matures, it will only be increasingly capable of unfairly competing with small creators. The livelihoods of these hardworking artists, writers, and musicians are particularly vulnerable to the AI revolution, and Congress can step in to protect them from unfair impacts. As AI moves forward, we must ensure that small creators are not left behind.

Background: Generative Artificial Intelligence (GenAI)

In recent months, Congressional interest in AI has exploded. AI is a broad term encompassing many kinds of software, each with profound implications for the sectors they touch. As such, the policy discussion surrounding AI and potential regulation and oversight is rightfully just as broad. However, the scope of this staff report is much narrower: it treats the role of copyrighted content in training GenAI, a subset of AI technology, and highlights the related concerns of America's small creators, many of whom are the sole proprietors of their businesses.

Overview

GenAI takes user instructions to create content such as text, images, videos, or audio.[1] It does this by combining elements of the target medium in configurations that are statistically likely to fit into the given parameters.[2] To make works that adequately fulfill user requests, GenAI models are trained on datasets to learn how to categorize, compare, and associate source materials, and notice patterns or differences within. In the same way that human creators can learn by observing others' work, GenAI models become capable of creating items like those included in their training data.[3]

GenAI models are known for their ability to adapt, combine, and prioritize information in their creations, creating work that can appear human-made. They can also frame information in unconventional ways, providing new perspectives and adding value to human-made work and thinking.[4]

The most notable introduction of GenAI in the technology scene happened in 2020 when OpenAI released GPT-3,[5] a textual GenAI model, the largest GenAI model at that time. Subsequently, competitors released their own models, and OpenAI then revealed ChatGPT, an improved version of GPT-3. While they received large amounts of public attention and scrutiny, they are not sudden developments. These models are built upon other, already existing technologies and are maturing more quickly because of increasing computing power and access to training data.[6]

---

[1] CONG. RESEARCH SERV., GENERATIVE ARTIFICIAL INTELLIGENCE AND COPYRIGHT LAW, 1 (May 11, 2023).

[2] Chris Stokel-Walker & Richard Van Noorden, *What ChatGPT and Generative AI Mean for Science*, NATURE (Feb. 6, 2023), https://www.nature.com/articles/d41586-023-00340-6; Kim Martineau, *What is Generative AI?*, INT'L BUS. MACHS. CORP., (Apr. 20, 2023), https://research.ibm.com/blog/what-is-generative-AI.

[3] GOV'T ACCOUNTABILITY OFFICE, SCIENCE & TECH SPOTLIGHT: GENERATIVE AI, 1 (Jun. 13, 2023); CONG. RESEARCH SERV., GENERATIVE ARTIFICIAL INTELLIGENCE AND DATA PRIVACY: A PRIMER, 2 (May 23, 2023); Ziv Epstein et al., *Art and the Science of Generative AI: A Deeper Dive,* 3 (Jun. 7, 2023), https://arxiv.org/abs/2306.04141.

[4] Tanay Varshney & Annie Surla, *An Introduction to Large Language Models: Prompt Engineering and P-Tuning*, NVIDIA CORP. (Apr. 26, 2023), https://developer.nvidia.com/blog/an-introduction-to-large-language-models-prompt-engineering-and-p-tuning.

[5] The term GPT is short for general-purpose transformer. In this context, a transformer is a kind of neural network (see "Underlying Technologies") that can understand the meaning of entire sentences, rather than individual words, and recall information previously stated in a given conversation. According to the Congressional Research Service, the creation of transformers and GPTs represented a leap in GenAI technology and paved the way to its mainstream use.

[6] Stokel-Walker & Van Noorden, *supra* note 2; Jim Euchner, *Generative AI*, 66 RES.-TECH. MGMT. 71 (2013), https://www.tandfonline.com/doi/full/10.1080/08956308.2023.2188861; Annamalai Chockalingam et al., *A Beginner's Guide to Large Language Models: Part 1*, 23, NVIDIA CORP. (2023), https://resources.nvidia.com/en-us-

Underlying Technologies

Machine Learning

Machine learning algorithms use given, correctly labeled content to statistically generate predictions and label new content. A more advanced iteration of machine learning, deep learning, can be given unlabeled, disorganized content, automatically label it, and continue to make predictions and other labels.[7]

Neural Networks

A neural network is a sophisticated type of deep learning algorithm that can identify and infer patterns in content.[8] Based on the human brain, it is structured into layers of nodes (i.e., computer connections) that take and create information and can "learn" from examples and instruction, just like the human brain can, without much or any human supervision.[9] In the same way that human brains are more powerful when they have more connections between brain cells, neural networks are quicker and more accurate when they have more nodes and layers. GenAI models use neural networks to understand the meaning of source materials, which materials are similar, why they are similar, and how to make new materials just like the sources.[10]

Foundation Models

A foundation model is a kind of neural network that has been pre-trained on a specific medium (e.g., text or images) with many examples in its dataset and can be adapted for various tasks. AI firms build their GenAI models, among others, upon foundation models, thus giving them their name.[11]

Capabilities

Using elements of examples in their training datasets, GenAI software creates content in their target media that is most likely to fit the instructions in given prompts.

---

large-language-model-ebooks/llm-ebook-part1; CONG. RESEARCH SERV., ARTIFICIAL INTELLIGENCE: BACKGROUND, SELECTED ISSUES, AND POLICY CONSIDERATIONS, 1 (May 19, 2021).

[7] Reinaldo Padilha França et al., *An Overview of Deep Learning in Big Data, Image, and Signal Processing in the Modern Digital Age*, in TRENDS IN DEEP LEARNING METHODOLOGIES: ALGORITHMS, APPLICATIONS, AND SYSTEMS 63 (Vincenzo Piuri et al. eds., 2021); *What is Machine Learning?*, INT'L BUS. MACHS. CORP., https://www.ibm.com/topics/machine-learning.

[8] *What is a Neural Network?*, AMAZON WEB SERV., https://aws.amazon.com/what-is/neural-network.

[9] França et al., *supra* note 7; Chockalingam et al., *supra* note 6, at 12-3; Int' Bus. Machs. Corp., *supra* note 7.

[10] *What is Generative AI?*, NVIDIA CORP., https://www.nvidia.com/en-us/glossary/data-science/generative-ai; Int'l Bus. Machs. Corp., *supra* note 7.

[11] *A New Era of Generative AI for Everyone*¸ 3, ACCENTURE (2023), https://www.accenture.com/content/dam/accenture/final/accenture-com/document/Accenture-A-New-Era-of-Generative-AI-for-Everyone.pdf; Cong. Research Serv., *supra* note 3, at 6; Nvidia Corp., *supra* note 10.

GenAI Can Make Art

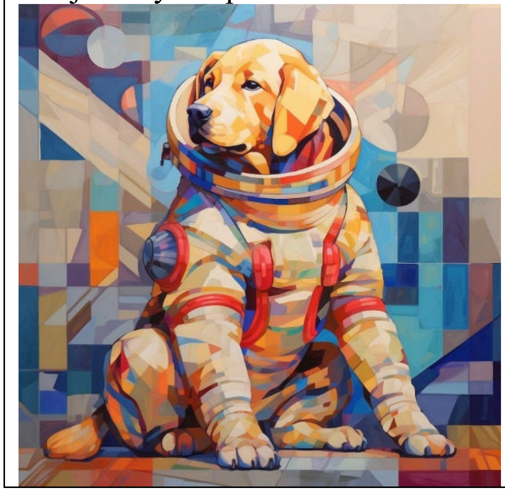
Social media platforms and political campaigns have been flooded with AI-generated images. Some have been realistic "deep fakes,"[12] while others are striking, new images using stylistic and creative elements found in existing art.[13]

When given a prompt, a GenAI art software generates images that are statistically like the images in its training data that it associates with the words or image exemplars in the prompt. This allows the software to create totally new content, combining elements of its training images and remaining faithful to given instructions.[14] The methodology also makes this software highly versatile.

Figure 1. GenAI Art Output Example 1[15]

Prompt: Imagine a Labrador dog in a space suit on the moon, Cubism.

Midjourney Output:



---

[12] CONG. RESEARCH SERV., DEEP FAKES AND NATIONAL SECURITY, 1 (Apr. 17, 2023).

[13] Kevin Roose, *An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy.*, N.Y. TIMES (Sep. 2, 2022), https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html.

[14] Gov't Accountability Office, *supra* note 3, at 1.

[15] Andris Mucenieks, *Midjourney Prompts 101 (With Examples),* PRINTIFY (Jun. 22, 2023), https://printify.com/blog/midjourney-prompts.

Figure 2. GenAI Art Output Example 2[16]

Prompt: Imagine a closeup shot of a woman playing the violin in an orchestra, focused and confident, shot on Fujifilm.

Midjourney Output:

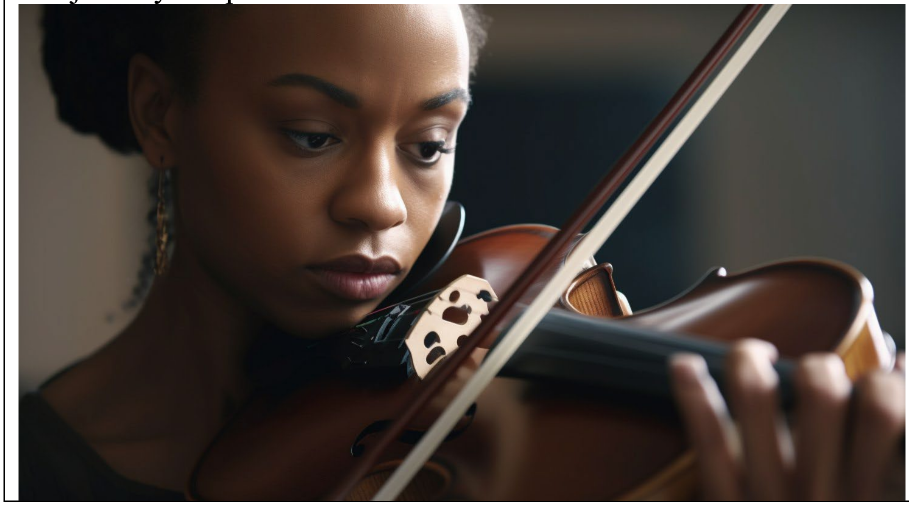
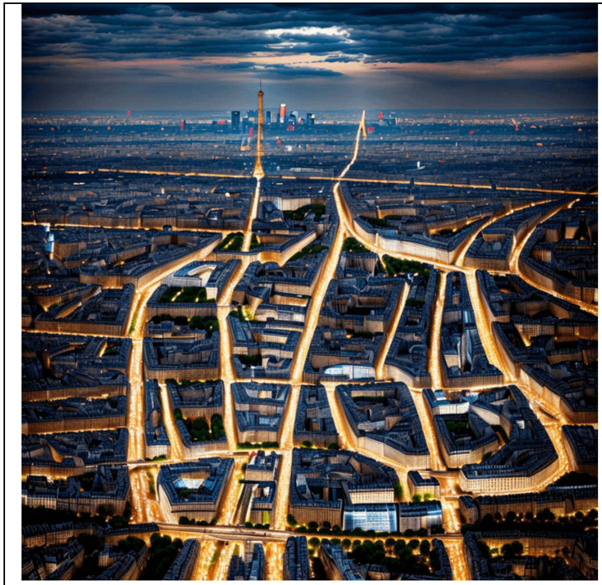
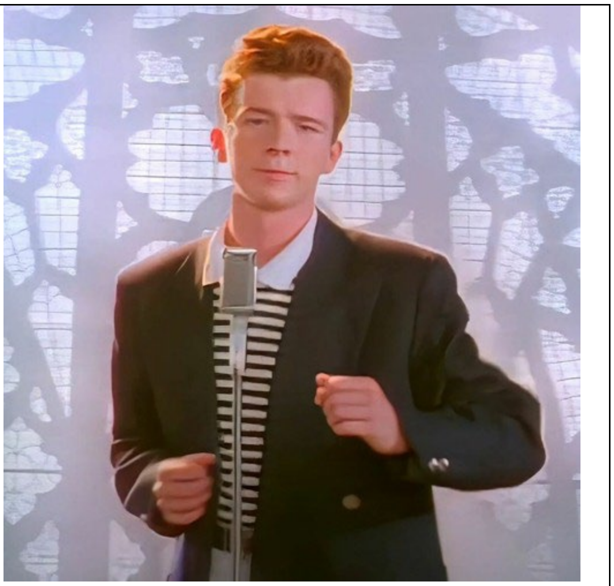


Figure 3. GenAI Art Output Example 3[17]

 *An AI-generated photorealistic image of a city's street grid which resembles a scene in Rick Astley's "Never Gonna Give You Up" music video.*	 *The corresponding scene from Rick Astley's "Never Gonna Give You Up" music video.*

---

[16] Nick St. Pierre (@nickfloats), Twitter (Mar. 15, 2023, 5:44 PM), https://twitter.com/nickfloats/status/1636122631614128134?s=20.

[17] *Hidden Imagery in AI Art – Never Gonna Give You Up*, KNOWYOURMEME, https://knowyourmeme.com/photos/2666059-hidden-imagery-in-ai-art (last visited Nov. 3, 2023); Rick Astley, *Never Gonna Give You Up (Official Music Video),* YOUTUBE (Oct. 25, 2009), https://www.youtube.com/watch?v=dQw4w9WgXcQ.

GenAI Can Creatively Write

The public has been highly engaged with GenAI text software like ChatGPT. Software like it can present information in a specific style or format, predict information, and combine, summarize, and simplify information.[18] The high use volume of platforms like ChatGPT has also led to their unforeseen, autonomous development of new capabilities like multiplication, interpreting combinations of emoji, and writing and running their own code.[19]

When given a prompt, GenAI text software combines words and text in a manner that is most likely to fit the instructions. Armed with millions or billions of points of training data and sample texts, this software is highly versatile and can write nearly any conceivable kind of text upon request. More advanced GenAI text software can also use and search for information on the Internet in real time while responding to prompts.[20]

Figure 4. GenAI Text Output Example

Prompt: Write a series of verses in the style of the Iliad, the Odyssey, and other Homeric works depicting the Small Business Administration's 7(a) loan guaranty program application process.

ChatGPT Output:
Sing, Muse, of the Small Business Administration's might,
And the 7(a) loan guaranty, a beacon of financial light.
A tale of entrepreneurs seeking funds for their grand dreams,
Through a process, intricate and complex, so it seems.

In the halls of government, the SBA doth reside,
Where business owners, with ambition, come to confide.
Their hopes and aspirations, they lay before the throne,
To obtain a loan to build, expand, or call their own.

The first step, like Achilles, in this heroic quest,
Is to prepare documentation, to prove your worthiness best.
A business plan, financials, and details of your trade,
Like a warrior's armor, must not be displayed.

Next, like Odysseus navigating the raging sea,
Applicants must seek a lender, the SBA's decree.
With a trusted partner, they shall set sail,
Hoping for approval, or their efforts shall fail.

Then the lender shall scrutinize, like a Cyclops' eye,

---

[18] CONG. RESEARCH SERV., GENERATIVE AI, 2 (May 23, 2023); Alex Hughes, *ChatGPT: Everything You Need to Know About OpenAI's GPT-4 Tool*, BBC SCI. FOCUS (Jun. 30, 2023), https://www.sciencefocus.com/future-technology/gpt-3.

[19] Stephen Ornes, *The Unpredictable Abilities Emerging from Large AI Models*, QUANTA MAGAZINE (Mar. 16, 2023), https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316.

[20] Gov't Accountability Office, *supra* note 3, at 1; Accenture, *supra* note 11, at 3; Cong. Research Serv., *supra* note 17, at 2; Alex Hughes; *supra* note 17.

The application, the credit, with judgment from on high.
To see if the borrower is worthy of the loan,
Or if their dreams of growth must forever postpone.

As if battling sirens, they must resist the urge,
To embellish their story or misrepresent the surge.
For the SBA's scrutiny is keen and astute,
And honesty's the path to approval, to boot.

Once past the lender's gate, like a journey's halfway mark,
The SBA shall deliberate, in chambers dim and dark.
With wisdom like Athena, they'll make the final call,
To grant or deny the loan, to rise or to fall.

Should fortune favor the bold, and the loan's approved,
The borrower shall embark, like a hero unmoved.
On a path paved with funds, to grow their enterprise,
With the SBA's support, they'll reach for the skies.

But if denied, they must persevere, like the heroes of old,
For another chance may come, like a fable retold.
With resolve and persistence, they'll try once more,
To secure their 7(a) loan, their business to restore.

So, sing, Muse, of the Small Business Administration's grace,
In aiding entrepreneurs, in this financial race.
A tale of courage, hope, and striving for success,
In the world of business, a true epic, no less.

GenAI Can Make Audio and Music

GenAI sound and music software can make content in the same way that GenAI art and text platforms can. By combining sound elements in a way most likely to fit given specifications and parameters, these tools can make highly realistic noises, verbal statements, and music. When creators use this software in tandem with GenAI text software, they can create content that almost perfectly mimics the style, voice, and unique tics of specific human musicians.[21]

There are two very common kinds of GenAI audio outputs that proliferate and go viral on the Internet. In the first type, users of GenAI voice cloning software create lifelike recordings of musical artists singing songs they have not actually sung, resulting in AI-generated covers of copyrighted material.[22] For instance, videos exist on YouTube nearly-authentically depicting Michael Jackson, who died in 2009 and did not speak Korean, singing the K-pop group BTS's 2020 hit "Dynamite."[23]

---

[21] Joe Coscarelli, *An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World*, N.Y. TIMES (Apr. 19, 2023), https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html.

[22] *Id.*

[23] MICHAEL JACKSON – DYNAMITE (AI COVER), (May 15, 2023), https://www.youtube.com/watch?v=C66HB6sDmo0 (last visited Jul. 28, 2023).

In the second type of GenAI music software output, users make all-new songs that appear to be written and recorded by artists and their producers. The most notable example of this content is an AI-generated recording depicting Drake and The Weeknd singing an original song, "Heart on My Sleeve," evoking their musical style, mirroring the topics they would sing about, and making relevant pop culture references. "Heart on My Sleeve" was even released on commercial music streaming platforms before it was removed due to the presence of a producer's protected mark in the music.[24] The software used to generate "Heart on My Sleeve" appeared to have been trained on so much copyrighted Drake and The Weeknd music that to fit any prompt asking for music that could have been made by both, the software deemed it necessary to include their shared producer's mark to make the output seem realistic.

GenAI Cannot Think for Itself Yet

Because GenAI makes content by stringing together statistically likely patterns of elements in the target medium, it is currently unable to reason and think critically. As such, GenAI software cannot yet fully understand the significance of elements of content. It can only understand how elements of content could be combined appropriately. For instance, GenAI text software like ChatGPT cannot fundamentally grasp why a joke is funny and make other jokes based on its understanding—it can merely only make sentences that sound like the (funny) jokes it has trained upon.[25]

This shortcoming is why GenAI developers attempt to program and train their models to refuse making certain outputs, like bigoted statements, misinformation, pornography, and other inherently harmful content.[26] Still, GenAI software can be easily jailbroken and made to create content from which developers aim to protect users.[27] GenAI text software can also include inaccurate information in its outputs[28] and even spontaneously make up false claims or facts, and assert they are real, an issue known as "hallucination."[29]

Even if GenAI software has been instructed to avoid or refrain from making specific kinds of content, it cannot do so in a foolproof way. Additionally, while GenAI software cannot specifically replicate copyrighted content in its training datasets, it can replicate elements of that content without fully understanding the implications of what it is doing and without the ability to fully refrain from replicating those elements.

---

[24] Joe Coscarelli, *supra* note 21.

[25] Cal Newport, *What Kind of Mind Does ChatGPT Have?*, THE NEW YORKER (Apr. 13, 2023), https://www.newyorker.com/science/annals-of-artificial-intelligence/what-kind-of-mind-does-chatgpt-have.

[26] Rebecca Klar, *AI Chatbots Provided Harmful Eating Disorder Content: Report,* THE HILL (Aug. 7, 2023), https://thehill.com/policy/technology/4141648-ai-chatbots-provided-harmful-eating-disorder-content-report; Billy Perrigo, *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic*, TIME (Jan. 18, 2023), https://time.com/6247678/openai-chatgpt-kenya-workers.

[27] Will Oremus, *The Clever Trick That Turns ChatGPT into Its Evil Twin*, WASH. POST (Feb. 14, 2023), https://www.washingtonpost.com/technology/2023/02/14/chatgpt-dan-jailbreak.

[28] Alex Janin, *What Happens When You Let a Chatbot Plan Your Meals*, WALL ST. J. (Aug. 14, 2023), https://www.wsj.com/articles/chatgpt-artificial-intelligence-recipes-meal-planning-edcd4a71.

[29] Karen Weise & Cade Metz, *When A.I. Chatbots Hallucinate*, N.Y. TIMES (May 1, 2023), https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html.

Training Data

Aside from the vast amounts of energy, money, and skilled manpower[30] needed to create a GenAI software, and build its underlying systems, a GenAI software's most important resource is training data. Any content in the target medium can be used as training data. This criterion is broad because every kind of GenAI software needs to examine and label examples of the content it is intended to make.[31] That way, it can understand what to generate and how to make outputs realistically fit user requests. For instance, to know how to make pictures of cats, a GenAI image software, like Midjourney, would need to know what cats look like, as well as the visual elements that make a cat a cat, and therefore examine existing pictures of cats in its training dataset.[32]

GenAI is Trained on Highly Expansive Datasets

To be highly flexible and capable of making almost any conceivable output, GenAI software must be trained on — in technical terms — a "large" dataset.[33] That said, the term "large," from a layperson's perspective, grossly understates the size of a typical GenAI training dataset: most datasets are voluminous archives consisting of millions, or even billions, of content examples.[34] In the context of digital storage, these archives take up dozens of terabytes: in fact, estimates pin the size of GPT-3's training dataset at 45 terabytes.[35]

The truly massive size of GenAI training datasets is crucial to their versatility: general-purpose datasets need to include content related to every conceivable prompt so that the software trained on those datasets can respond to all kinds of user prompts. Conversely, if a GenAI software was intended to make niche content, create very specific kinds of work, or use standardized formatting, its dataset would need to be narrowed and fine-tuned.[36] For example, to write the AI-generated text example displayed in Figure 4, ChatGPT needed to have many copies of information about the Small Business Administration (SBA)'s 7(a) loan guaranty program and existing Homeric works stored in its training dataset to incorporate and emulate. In other words, for a GenAI software to emulate *anything*, it must "understand" *everything*.

---

[30] Nvidia Corp., *supra* note 10; *What is Generative AI?*, McKinsey & Co. (Jan. 19, 2023), https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai#; Kathy Baxter & Yoav Schlesinger, *Managing the Risks of Generative AI*, Harv. Bus. Rev. (Jun. 6, 2023), https://hbr.org/2023/06/managing-the-risks-of-generative-ai.

[31] Cong. Research Serv., *supra* note 1, at 3; Cong. Research Serv., *supra* note 3, at 3; Sheera Frankel & Stuart A. Thompson, *'Not for Machines to Harvest': Data Revolts Break Out Against A.I.*, N.Y. Times (Jul. 15, 2023), https://www.nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html.

[32] Gov't Accountability Office, *supra* note 3, at 1.

[33] The characterization "large" also appears in the term "large" language model (LLM), which is a type of foundation model trained specifically on text and written language. ChatGPT and other GenAI text software are built upon LLMs.

[34] Gov't Accountability Office, *supra* note 3, at 1; Cong. Research Serv., *supra* note 3, at 3-4; Annamalai Chockalingam et al., *supra* note 6, at 8-11; Accenture, *supra* note 11, at 3.

[35] McKinsey & Co., *supra* note 30.

[36] Gov't Accountability Office, *supra* note 3, at 1; Nvidia Corp., *supra* note 4; Nvidia Corp., *supra* note 6, at 10; Accenture, *supra* note 11, at 6.

The need to have an immense volume of content, and a wide variety of it, makes assembling training datasets a daunting, expensive, and cumbersome task.[37] To alleviate licensing costs and the menial tasks involved in finding and adding content to the dataset, developers often turn to easily accessible content taken from the Internet by automated software. However, much of this content, though easy to obtain, may be protected by their creators' intellectual property rights, pending court litigation or legislative action.[38]

GenAI Outputs Are Significantly Impacted by Items Included in Training Datasets

As discussed in the previous section, GenAI models need exemplars of the content they aim to create for users in their training datasets. This makes it highly plausible that including specific content in a GenAI model's dataset would significantly impact its outputs.

For example, in a recent copyright and trademark infringement lawsuit, Getty Images alleges that the GenAI image software Stable Diffusion mimics Getty watermarks in its outputs because photographs with Getty watermarks – showing that they are copyrighted and owned by Getty – appear in the software's training data. Without those photographs' presence in the training data, Stable Diffusion would be unable to emulate Getty watermarks in its own outputs.[39] This logic similarly holds in the SBA 7(a) loan guaranty program Homeric verse example mentioned in the previous section: without an example of a Homeric verse and an example of SBA 7(a) program information in its training data, ChatGPT would be unable to understand what each was, or how to emulate them.

Furthermore, the unconscious bias inherent in much of the Internet's content is reflected in many GenAI outputs. Documentation exists of cases where GenAI image software has distorted, stereotyped, and censored depictions of Black people, history, and culture in its outputs. These incidents, critics contend, occur because stereotypes and prejudiced content regarding marginalized groups have been overrepresented in some GenAI training datasets. That said, GenAI firms such as OpenAI, Midjourney, and Stability AI (the company behind Stable Diffusion) have commendably acknowledged the issue of bias in their training data and pledged to improve their tools.[40]

This mechanic also highlights the importance of marking undesirable kinds of content so that a GenAI model can refuse to create similar outputs.[41] Still, as mentioned in the section "GenAI Cannot Think for Itself Yet," methods currently exist to jailbreak GenAI models and force them to make harmful content—drawing upon the harmful content that is included in their training data.

---

[37] CONG. RESEARCH SERV., ARTIFICIAL INTELLIGENCE: OVERVIEW, RECENT ADVANCES, AND CONSIDERATIONS FOR THE 118TH CONGRESS, 2-3 (Aug. 4, 2023).

[38] *Id,* at 3; Cong. Research Serv., *supra* note 1, at 4.

[39] *See* Complaint at 18, Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135 (D.Del. Feb. 3, 2023), https://copyrightlately.com/pdfviewer/getty-images-v-stability-ai-complaint.

[40] Zachary Small, *Black Artists Say A.I. Shows Bias, With Algorithms Erasing Their History*, N.Y. TIMES (Jul. 4, 2023), https://www.nytimes.com/2023/07/04/arts/design/black-artists-bias-ai.html.

[41] Billy Perrigo, *supra* note 26.

All of this is to underscore the integral part that training data plays in the quality and outputs of GenAI models. As many observers write, GenAI is only as good as the data upon which it is trained.[42]

"Black Box" Issue: GenAI Decision-Making and Training Processes Are Nontransparent

Even with the recent attention on GenAI and its training data, not much is publicly known about exact dataset details or the way GenAI software makes decisions, as developers do not often disclose them. Because developers do not share how their GenAI models work, it becomes extremely difficult for researchers, policymakers, and the public to identify and remedy the issues and factors underlying the choices these models make.[43] Additionally, developers do not even fully understand some of the emergent, unexpected capabilities that their GenAI models develop, or how they came into being.[44] This set of problems is commonly referred to as the "black box" issue: GenAI software is considered a "black box" because observers cannot see inside the opaque processes that turn inputs into outputs.[45]

To address the "black box" issue, much discussion has focused on transparency and explainability, core tenets of Senate Majority Leader Chuck Schumer's recently proposed *SAFE Innovation Framework*.[46] Furthermore, explainability and understanding emergent capabilities in AI, which was previously overlooked, is now a topic of great interest in the academic community. However, the "black box" issue is complicated and cannot be resolved overnight. Much thought needs to be put into striking the right balance between transparency and encouraging private innovation in AI.

GenAI Has Serious Ramifications for Small Creators

As repeated ad nauseam in the media and policy conversations surrounding AI, this technology is very likely to fundamentally remake how society approaches work and creativity. In the realm of content creation, GenAI can automate many aspects of content creation and acquisition, making it vastly quicker, cheaper, and easier for users compared to manual, human work.

GenAI Competes with Human Creators and Devalues Their Work

Because many GenAI outputs are nearly indistinguishable from human-made content, and so quick and cheap to obtain,[47] human content creators are increasingly finding themselves competing against GenAI firms in securing work and revenue.[48] Even though GenAI isn't yet completely

---

[42] Kathy Baxter & Yoav Schlesinger, *supra* note 30; Dana Rao, *Responsible Innovation in the Age of Generative AI*, ADOBE (Mar. 21, 2023), https://blog.adobe.com/en/publish/2023/03/21/responsible-innovation-age-of-generative-ai.

[43] Cong. Research Serv., *supra* note 1, at 4; Cong. Research Serv., *supra* note 6, at 12; Ziv Epstein et al., *supra* note 3, at 6.

[44] Cong. Research Serv., *supra* note 37, at 2.

[45] CONG. RESEARCH SERV., ARTIFICIAL INTELLIGENCE (AI) AND EDUCATION, 2 (Aug. 1, 2018).

[46] *Sen. Chuck Schumer Launches SAFE Innovation in the AI Age at CSIS*, 7-8, CTR. FOR STRATEGIC AND INT'L STUDIES (Jun. 21, 2023), https://csis-website-prod.s3.amazonaws.com/s3fs-public/2023-06/230621_Schumer_SAFE_Innovation.pdf?VersionId=jApHm2QrP7nAZvl_B4GJ6s_YjSrfyYBK.

[47] Ziv Epstein et al., *supra* note 3, at 9.

[48] McKinsey & Co., *supra* note 30; Alain Strowel, *ChatGPT and Generative AI Tools: Theft of Intellectual Labor?*, 54 IIC 491, 492, (2023), https://link.springer.com/10.1007/s40319-023-01321-y; Joel Loynds, *Adobe Firefly Burns*

mature — there are still notable issues in the way GenAI image software portrays people of color[49] — its versatility and speed give it an objectively inhuman advantage over its human competitors. GenAI software can quickly give users outputs that fit their specifications for much lower prices than commissions would cost for human creators.[50] The extremely high profile and interest GenAI software has received in recent months has also made it very easy to find, as opposed to many small creators who struggle to gain exposure in the content market.

Such developments could lead to the concentration of market share in both the content creation and AI markets in the hands of a few GenAI firms and disincentivize human content creation.[51] In fact, reports already state that companies and other potential clients for content creators are instead turning to GenAI firms to generate visual content for social media outreach.[52] Without adequate guardrails, the proliferation and increasing availability of GenAI software could fundamentally threaten the livelihoods of many of America's small creators.

GenAI Uses Human Creations to Learn and Mimic, Unbalancing Competition with Human Creators

As previously mentioned, GenAI relies on the work of its human competitors to function and improve. Its ability to quickly assimilate the skills and aesthetic styles of others through the inclusion of content in training data,[53] is much faster than the years of painstaking training that humans undergo in learning their crafts. Armed with training data containing recent, groundbreaking content, GenAI software can suddenly enter the market for that novel, niche content style, with experience in every other conceivable style.

Such a dynamic could further depress human content creation and sharing. As more and more human creators find their work being incorporated into competing GenAI training datasets, they may choose to stop sharing their content and work, staunching the free flow of ideas and inspiration and straining the digital commons.[54] The growing knowledge that work shared online can be incorporated into GenAI training datasets without compensation could spook human creators from posting their work or sharing samples of their content, further harming their advertising opportunities and their prospects of finding work.

The sheer inherent advantages that GenAI models have over human creators in speed, ubiquitous name recognition, and versatility make the playing field extremely unlevel when it comes to competition between humans and GenAI.

Bridges with Stock Photo Creators, Dexerto (Jun. 21, 2023, https://www.dexerto.com/tech/adobe-firefly-burns-bridges-with-stock-photo-creators-2185518.

[49] Zachary Small, *supra* note 40.

[50] Saffron Huang & Divya Siddarth, *Generative AI and the Digital Commons*, 4-5, (Mar. 20, 2023), http://arxiv.org/abs/2303.11074.

[51] *Id*, at 2-5.

[52] Accenture, *supra* note 11, at 7.

[53] Ziv Epstein et al., *supra* note 3, at 3.

[54] Joel Loynds, *supra* note 48; Saffron Huang & Divya Siddarth, *supra* note 50, at 7.

“Black Box” Issue Obscures Potential Harms and Risks

It is also important to note that the “black box” issue can make detecting these risks and harms difficult, if not impossible, for every actor in the AI policy space. GenAI firms often do not disclose what is included in their training datasets, which means that they do not guarantee small creators that their work has not been included.[55] Furthermore, academics, policymakers, and even developers themselves cannot explain precisely how the inclusion of a given piece of content in a GenAI training dataset would affect its outputs, the impacts of its removal, and how the accompanying model picks elements from its training data to mimic. While it is easy to understand that America’s small creators are harmed by the use of their work without compensation in competing GenAI models’ training, it is hard to determine the extent of the damage without a better understanding of the processes that underlie these models’ training and “creative” decision-making.[56]

Recently, it was revealed that one GenAI firm is under investigation by the Federal Trade Commission to determine in part how its models are trained and potentially where it sources its training data. The firm, OpenAI, the maker of ChatGPT, is believed to be reluctant to share such information because of the resulting undesirable opportunity for competitors to copy its methods.[57] While this concern is understandable, it should not be necessary for a federal agency to investigate a firm to force transparency in a potentially industry-wide practice that could harm a vast swath of America’s small businesses.

Ongoing Litigation and Need for Congressional Action

As referenced previously, ongoing litigation could change the way the law treats GenAI and its use of human-made content for training.[58] However, it is important to note that these legal actions, while capable of remedying the harm done by these GenAI training practices, could take too long for many of America’s small creators to survive on their own. Furthermore, long, drawn-out lawsuits, as a practical matter, can only be sustained by corporations and the wealthy. America’s small creators cannot afford to initiate and fight their own lawsuits to obtain relief and compensation. Congress must act in upcoming AI legislation to address this issue and maximize benefits and protections for American small creators.

Policy Considerations

The conversation surrounding GenAI policy has drawn much interest and many perspectives. This report treats four of the most concrete and discussed policy options to address this issue: a statutory mandate for GenAI firms to specifically disclose the presence of copyrighted content in their training datasets; a licensing, opt-out, and compensation scheme for human-made content that

---

[55] *Id*, at 13; Ziv Epstein et al., *supra* note 3, at 6.

[56] Cong. Research Serv., *supra* note 6, at 27; Ziv Epstein et al., *supra* note 3, at 4.

[57] Cecilia Kang & Cade Metz, *F.T.C. Opens Investigation into ChatGPT Maker Over Technology’s Potential Harms*, N.Y. TIMES (Jul. 13, 2023), https://www.nytimes.com/2023/07/13/technology/chatgpt-investigation-ftc-openai.html.

[58] Cong. Research Serv., *supra* note 1, at 4; Judy Kurtz, *Sarah Silverman Sues OpenAI, Meta Over Use of her Book in Training AI Programs*, THE HILL (Jul. 10, 2023), https://thehill.com/blogs/in-the-know/4088765-sarah-silverman-sues-openai-meta-over-use-of-her-book-in-training-ai-programs.

could be used to train a GenAI model; the creation of a federal, public "gold standard" training dataset with properly licensed content for the use of GenAI developers; and a watermarking scheme for GenAI outputs.

Mandated Disclosure of Unlicensed and Copyrighted Content in Training Datasets

It is believed that members of the European Union Parliament, in writing regulations for AI, plan to require software like GenAI to disclose the presence of unlicensed and copyrighted content in training data. While controversial, the adoption of similar transparency requirements at the federal level would address the "black box" issue and make it easier for small creators to determine whether their work has been used for GenAI training without their consent.

GenAI firms may raise concerns that revealing even parts of their training datasets could allow competitors to copy their training data. It is important to remember that to assemble these datasets in the first place, a firm needs large amounts of funding and other resources. With these resources, it is likely that these firms can afford to resolve these disputes among themselves. Furthermore, a leaked internal memo from Google argues that many of these closed-source GenAI models "have no secret sauce" and are rapidly being outpaced by open-source models,[59] meaning that the firms who embrace transparency in their training data may not necessarily lose advantages against closed-source GenAI developers.

Mandating GenAI firms' disclosure of unlicensed and copyrighted data in training datasets would allow creators to more easily determine whether their work has been used without their consent and the extent to which that potentially unfair use hurts their businesses. The added transparency would help present a check against the harm that indiscriminate content scraping poses to the livelihoods of America's small creators.

Licensing and Compensation for Use of Content in Training Data

Some policy actors have called for the creation of a content licensing scheme for GenAI training data collection, to be used in the same way that commercial entities can license copyrighted content and other protected intellectual property from others.[60] Stakeholders differ on how a licensing scheme should be implemented through statute, specifically whether it should run on government-set rates, if rates should be set on the free market, or if rates should be negotiated through artist collectives. Still, the idea, in general, would lead to the generation of revenue for creators whose work has been used for GenAI training while still giving GenAI firms a way to collect content created by others.

---

[59] Dylan Patel & Afzal Ahmad, *Google "We Have No Moat, and Neither Does OpenAI:" Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI*, SEMIANALYSIS (May 4, 2023), https://www.semianalysis.com/p/google-we-have-no-moat-and-neither; Emma Roth, *That Google Memo About Having 'No Moat' in AI Was Real – and Google's AI Boss Disagrees with It*, THE VERGE (Jul. 10, 2023), https://www.theverge.com/2023/7/10/23790132/google-memo-moat-ai-leak-demis-hassabis.

[60] Saffron Huang & Divya Siddarth, *supra* note 50, at 9; *Protect Working Musicians Act (PWMA) of 2023*, AM. ASS'N OF INDEP. MUSIC (Sep. 25, 2023), https://a2im.org/protect-working-musicians-act-2023/.

Opt-Out Systems and Do-Not-Train Tags

Many licensing proposals are also mentioned alongside proposals for government-backed standards and requirements for opt-out and do-not-train tags. Some GenAI firms have begun developing ways to mark online content so that their training datasets exclude marked websites and other content.[61] Allowing creators to explicitly identify their work as not licensed for GenAI training would make it easier for them to seek recourse if their work has been used without their consent, and for GenAI developers to understand what content to avoid when building training datasets. However, it is also important to ensure that whatever system is implemented is unified, reliable, and easy-to-use, so that small creators can focus on their work rather than navigating a complicated maze of opt-out and do-not-train systems that vary in scope and efficacy.

Clarifying Fair Use Exceptions

As previously mentioned, litigation focusing on GenAI training datasets' unlicensed inclusion of human-made content remains pending. These lawsuits hinge upon how federal copyright law is applied to such use. By clarifying in future legislation that fair use exceptions in copyright law do not apply to inclusion in GenAI training, Congress could effectively render GenAI firms' legal defenses relying upon fair use exceptions useless.[62] This change will make GenAI developers more vulnerable to legal action if they continue using human-made work without authorization. However, as also written above, many small creators do not have the means or resources to initiate and sustain lawsuits against firms as well-resourced as those capable of creating GenAI models. The direct, practical benefits of tightening copyright law would be limited, though not undesirable, for America's small creators.

"Gold Standard" Training Datasets

Stakeholders have suggested that an entity like the National Institute of Standards and Technology or the new National AI Research Resource create a "gold standard" training dataset with licensed and legally safe content available for public and commercial use in training GenAI models.[63]

Like the licensing proposal, implementing this idea would create another possible source of revenue for small creators while still providing GenAI developers with viable training data. Additionally, the creation of a "gold standard" dataset could have the added benefit of lowering barriers to entry into the GenAI market for new AI startup businesses. The creation of this standardized dataset could also demonstrate its viability and help nudge private companies to use

---

[61] Meera Navlakha, *OpenAI Launches Webcrawler GPTBot, and Instructions on How to Block It*, MASHABLE (Aug. 8, 2023), https://mashable.com/article/open-ai-gptbot-crawler-block; Forrester, *Generative AI Gets an Upgrade to Business Class*, FORBES (Mar. 31, 2023), https://www.forbes.com/sites/forrester/2023/03/31/generative-ai-gets-an-upgrade-to-business-class; Artificial Intelligence and Intellectual Property – Part II: Copyright: Hearing Before the Subcomm. on Intellectual Property, 118th Cong. (2023) (Statement of Dana Rao) [*hereinafter* Rao Judiciary Testimony].

[62] Cong. Research Serv., *supra* note 1, at 4; Judy Kurtz, *supra* note 58.

[63] Saffron Huang & Divya Siddarth, *supra* note 50, at 13.

licensed, legally safe training datasets. In fact, some firms, such as Adobe and Getty, have already created GenAI image tools trained only on properly licensed content.[64]

However, the creation of a federal government GenAI training dataset would raise numerous controversial questions about its nature, especially because of ongoing policy discussions surrounding misinformation and censorship. In funding or mandating the creation of such a dataset, it is imperative that Congress prohibit the inclusion of false, misleading, and other harmful content under the disingenuous guise of including the views of "both sides" of an ideological conflict such as the January 6, 2021 insurrection at the United States Capitol. It is important that any "gold standard" legislative proposal resolve these questions before advancing.

Watermarking GenAI Creations

Congress is considering legislation to require the watermarking of GenAI outputs, and President Biden's Executive Order 14,110 directs the Commerce Department to report on "science-backed standards and techniques" for GenAI output watermarking. GenAI firms are also exploring watermarking and content "nutrition label" -like technology.[65] These watermarks can be visual or digital, depending on the output medium, and technology is being developed that would make them difficult, if not impossible, to remove.[66] For optimal efficacy, the content ecosystem would need to widely adopt output watermarking with a single, universal standard[67] — something that can be prompted by federal regulation. This could be implemented through the creation of a standard watermarking system by an agency like the National Institute of Standards and Technology and a statutory requirement for GenAI software to watermark outputs.

While many conversations surrounding GenAI watermarking have focused mainly on deepfakes and misinformation, a policy mandating the practice could also be beneficial to small creators by distinguishing potentially competing content and helping potential clients and viewers understand what content has been made by a GenAI model and what has been made by a human. This distinction would be analogous to the same distinctions made between products made by American small businesses and products made abroad or products made by large corporations and would help inform client decisions in who to solicit content from. In promoting the use of human-made content, this kind of federal regulation could be a boon to America's small creators.

---

[64] Rao Judiciary Testimony, *supra* note 61; Press Release, Getty Images, Getty Images Launches Commercially Safe Generative AI Offering (Sep. 25, 2023), https://newsroom.gettyimages.com/en/getty-images/getty-images-launches-commercially-safe-generative-ai-offering.

[65] Rao Judiciary Testimony, *supra* note 61; Cristiano Lima, *Congress is Playing Catch-Up on AI. She's Been Sounding the Alarm,* WASH. POST (Jul. 13, 2023), https://www.washingtonpost.com/politics/2023/07/13/congress-is-playing-catch-up-ai-shes-been-sounding-alarm; Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Oct. 30, 2023).

[66] Ziv Epstein et al., *supra* note 3, at 9-10; John Kirchenbauer et al., *A Watermark for Large Language Models*, (Jun. 6, 2023), http://arxiv.org/abs/2301.10226; Keith Collins, *How ChatGPT Could Embed a 'Watermark' in the Text It Generates*, N.Y. TIMES (Feb. 17, 2023), https://www.nytimes.com/interactive/2023/02/17/business/ai-text-detection.html.

[67] Ziv Epstein et al., *supra* note 3, at 9-10.

Conclusion: GenAI Policy Must Benefit America's Most Vulnerable Content Creators

GenAI, in general, is not yet a fully mature technology. Many of its outputs still contain inaccuracies, distortions, and "glitches" that render them imperfect and flawed. However, even in its current state, GenAI software in various mediums has demonstrated that it can unfairly outcompete America's small artists, writers, and musicians, and push them out of business. Adding to the advantages that GenAI holds in content creation is the fact that GenAI firms regularly use these creators' work to train and improve their models, commercially benefitting at creators' expense. As GenAI continues to mature, its competition with human content creators will only intensify, and its advantages will grow wider.

Much fanfare has been made of the forthcoming action and debate surrounding AI policy and regulation, with visions of far-reaching, long-term legislation.[68] While it is important to adopt a far-sighted, comprehensive approach to addressing the AI revolution and the issues it will surface, it is also imperative that Congress directly acts to protect America's most vulnerable entrepreneurs and content creators from the shocks and consolidated corporate power that AI will bring. GenAI may be merely a part of the larger conversation surrounding AI, but it is extremely important to the small creators that the Committee endeavors to empower.

---

[68] Ctr. for Strategic and Int'l Studies, *supra* note 46; Ryan Tarinelli, *Lawmakers Suggest Agency to Supervise Artificial Intelligence,* ROLL CALL (May 23, 2023), https://rollcall.com/2023/05/23/lawmakers-suggest-agency-to-supervise-artificial-intelligence.