BEFORE THE
WORLD INTELLECTUAL PROPERTY ORGANIZATION

In re Impact of Artificial Intelligence on Intellectual Property Policy	WIPO/IP/AI/2/GE/20/1

Comments of Engine Advocacy

Engine is a non-profit technology policy, research, and advocacy organization that bridges the gap between policymakers and startups. Engine works with government and a community of thousands of high-technology, growth-oriented startups across the nation to support the development of technology entrepreneurship. We appreciate the World Intellectual Property Organization's ("WIPO") attention to pursuing policies that foster innovation. And we further appreciate the opportunity to submit these comments regarding the intersection of artificial intelligence ("AI") technologies and intellectual property ("IP") law and policy, as well as the impact IP law and policy could have on emerging AI technology.

AI does not need to disrupt IP law and policy, and in our view current U.S. IP law and policy frameworks are working well. The following comments address four of the issues raised in WIPO's Draft Issue Paper which are particularly pertinent to innovators today. Engine previously submitted comments to the U.S. Patent and Trademark Office ("USPTO") on the role of IP in AI technology, and our earlier comments (attached as Appendices A & B) are also relevant to the issues WIPO has raised.

A. Issue 2: Patentable Subject Matter and Patentability Guidelines

The law should prevent individuals and companies from obtaining patents that generically claim the performance of abstract ideas using conventional AI technology. For example, patentees should not be allowed to claim (and exclude others from) the use of off-the-shelf AI technology to perform a typical human mental task. Without more, that is not the type of inventive contribution warranting patent protection. And allowing that sort of overbroad claiming directed to abstract ideas would prevent downstream innovation and allow patentees to hold-up large sectors of productive economic activity by asserting or threatening to assert such overbroad patents.

AI inventions do not require a unique eligibility framework, and instead the current standards that apply to eligibility of other technologies should likewise be applied in the context of AI. Under U.S. law, patent subject matter eligibility is governed by 35 U.S.C. § 101 and assessed consistent with the framework articulated in *Alice Corp. v. CLS Bank International*.[1] Briefly, when considering the eligibility of an AI patent (like any patent), courts and patent examiners should determine whether the claims at issue are directed to an abstract idea.[2] If the claims are directed to an abstract idea, then the courts and examiners should consider whether the claims articulate an "inventive concept." Put another way, courts and examiners look at whether there are additional elements of the claim that, when considered individually or as an ordered combination, are "sufficient to ensure that the patent in practice amounts to significantly more than a patent upon the [abstract idea] itself."[3] If a claim is directed to an abstract idea and lacks that inventive concept, i.e., lacks that "something more," then it should fall outside the scope of eligibility. By contrast, if an AI patent claims a technology-based solution to a technological problem, then it is more likely to be patent eligible. For example, a patent may properly claim specific improvements to the performance of an AI system itself, and disclose a way to overcome existing problems with AI systems.[4] For a more detailed discussion on this issue, please refer to Appendix A, part 5 (pages 6-8).

B. Issue 4: Disclosure

We agree that sufficient disclosure is a fundamental feature of the patent system, and in some ways the standards that ensure proper disclosure are even more important—and should be applied rigorously—in the context of certain AI inventions. AI inventions inherently involve some level of unpredictability. For example, as we have previously explained, with AI systems "[t]here is no guarantee running the same algorithm on the same dataset will . . . reach the same result."[5] But it is still essential that those seeking AI patents provide a sufficient disclosure so that the public both (a) is on notice of what is (and is not) claimed and (b) has enough information to replicate (or practice) what is claimed. Otherwise, persons of ordinary skill will need to engage in lengthy trial-and-error to practice the patent and may have a difficult time understanding whether their independent research and development runs afoul of any patent claims.

At a minimum, those seeking AI patents should be required to disclose, in detail, both the model used and how to train it. This includes disclosing the model's architecture, the model's

[1] 573 U.S. 208 (2014).

[2] *Id.* at 217.

[3] *Id.* at 218 (quotations omitted).

[4] *See, e.g.*, *BASCOM Glob. Internet Servs., Inc. v. AT&T Mobility LLC*, 827 F.3d 1341, 1351 (Fed. Cir. 2016).

[5] Appendix A at 9.

objective function, how the AI model was trained, and the input data. And the disclosure of input data must be more than a generic description of data used to train an AI system. Instead, to satisfy enablement, a patentee could for example disclose his or her training data directly (or make it available through a public repository).[6] Alternatively, a patentee could include a detailed description of data in the patent specification, describing all data characteristics necessary for the AI to learn the tasks. For a more detailed discussion on this issue, please refer to Appendix A, Part 7 (pages 8-13).

C. Issue 7: Infringement and Exceptions

It should be lawful to ingest content and use it to train, tune, and/or test AI systems. And under current U.S. law, it is. As we have previously explained "[s]uch uses either involve content that is ineligible for copyright protection, involve uses that are not infringing, or involve fair uses of content."[7] To the extent future developments in AI render the current legal frameworks unduly ambiguous or unclear about the infringement liability associated with AI processing content, or if there are inconsistencies globally, then it may be appropriate to amend relevant laws to confirm that such uses are noninfringing.

If ingesting content to train, tune, and/or test AI systems were unlawful, that would open AI startups and innovators to unmanageable cost and risk and have a negative impact on the development of new technology or improvements thereto. As we have previously explained, "[c]hanging copyright law or policy in a way that creates liability when AI systems ingest content would put innovation at risk."[8]

For one, because AI is so pervasive and it relies on vast amounts of data, introducing new copyright liability now would have significant and far-reaching consequences. As we explained to the USPTO:

> AI systems are designed, and will continue to be designed, to solve many different types of problems in reliance on many different types of data. Similarly, content that is potentially eligible for copyright protection is ubiquitous and varied. Copyright protection [in the U.S.] can apply to *any* "original works of authorship fixed in [a] tangible medium of expression." While underlying facts, data, ideas, etc., are not eligible for copyright

[6] While this sort of direct, detailed disclosure of training data may be preferable from a patent policy perspective, it may not be possible. For example, if training data includes personally identifiable health or financial information that cannot not be disclosed without putting individual privacy at risk, the countervailing privacy interest would caution against full disclosure of training data in all AI patents.

[7] Appendix B at 2.

[8] Appendix B at 7.

> protection [in the U.S.], the broader works that contain underlying facts might have expressive elements that render such works copyrightable. As such, the data used to train, test, and tune AI systems may be taken from content that is potentially eligible for copyright protection.[9]

Relatedly,

> Everyone who generates any content arguably has some claim to copyright protection (and such content can range from highly expressive paintings to largely-factual academic papers that are at least expressive in part). It would be untenable to require innovators developing new AI technology to assess the copyright status of every piece of content and every data source feeding into an AI system. If developers then had to obtain licenses to any content or data sources, it would magnify the problem enormously. If developers faced those sorts of burdens when compiling datasets, it would slow (and could stall) progress.[10]

Not only would increasing copyright liability over AI uses of content put innovation at risk, but if it were more difficult for developers to navigate the copyright landscape when compiling datasets, it could result in them leaving out certain content. And that, in turn, could result in non-comprehensive datasets being used to train AI systems, and subjecting those systems to increased risk of bias. As we explained to the USPTO:

> Access to content free from claims of copyright infringement gives developers access to more diverse, less biased content. For example, recently created (and therefore currently copyrighted) works may be less susceptible to inherent or traditional biases, when compared to public domain content that was created when the U.S. was less racially diverse and had more gender disparity. And "[b]ias in AI may be exacerbated by a restrictive fair use doctrine," because if training data "is protected by copyright, those who use [it] do so secretly, preventing biases from being uncovered." Moreover, bias problems could arise if copyright owners were in a position to dictate when and who can develop AI technology and for which purposes.[11]

For a more detailed discussion on this issue, please refer to Appendix B.

[9] Appendix B at 2 (citations omitted).
[10] Appendix B at 8 (citations omitted).
[11] Appendix B at 7-8 (citations omitted).

D. Issue 9: General Policy Issues

Society is enjoying a rapid growth of AI technology, with new developments and advances emerging regularly and new commercial applications always on the horizon. All of that progress has occurred under the current IP frameworks, and policymakers should avoid making any changes that hamper or slow this innovation.

Likewise, policymakers should consider how any changes to current IP frameworks would affect the disclosure norms of the AI development community. There is already a high level of transparency and sharing among developers. AI systems and tools are often distributed freely through open source or permissive licenses. The patent system should naturally dovetail with these norms, because the disclosure requirements of patent law also promote this sort of information sharing. However, if AI patents start to issue without sufficient enabling disclosure, those patents would not only fail the law's requirements but would also undercut the open ideals of the developer community.

Finally, turning back to the risk of bias in AI: As noted above, AI systems will be less susceptible to bias if developers are allowed to continue to use content for training, tuning, and testing free from copyright liability. The alternative, of allowing individual rightsholders to decide what, when, and who can be involved in developing AI for what purposes, increases risks of bias. Likewise, in the patent context, opaque AI systems trained with biased datasets can negatively impact individuals and communities. The patent system itself can have a positive impact there. As we previously explained, "[s]trong patents with precise claims can better help the public understand how AI systems make decisions and impact their lives. And properly disclosing the data used to train AI systems in order to meet the enablement requirement can give the public the opportunity to check for bias in the dataset or data collection process."

For a more detailed discussion on this issue, please refer to Appendix A, Part 11 (pages 14-16) and Appendix B, Part V (pages 7-9).

E. Conclusion

Engine looks forward to continued engagement with WIPO and the USPTO on the intersection of AI technologies and IP law and policy. AI development is thriving, and recent advances in promising AI technology occurred under current IP frameworks. Before making changes, policymakers should carefully consider how any such changes would impact the trajectory of exciting AI technologies, and policymakers should avoid making changes that could stall innovation.

DATE: February 14, 2020

Abigail A. Rives
Intellectual Property Counsel
Engine Advocacy
700 Pennsylvania Avenue, SE
2nd Floor
Washington, DC 20003

Appendix A

Before the

U.S. Patent and Trademark Office

In the Matter of

Request for Comments on Patenting Artificial Intelligence Inventions

Docket Number PTO-C-2019-0029

Comments of Engine Advocacy & The Electronic Frontier Foundation

Tyler D. Robbins
Certified Law Student
Phil Malone
Counsel for Commenters

Juelsgaard Intellectual Property
& Innovation Clinic
Stanford Law School
(650) 724-1900
pmalone@stanford.edu

November 8, 2019

Contents

About the Commenters

Engine Advocacy – Engine Advocacy is a non-profit policy, research, and advocacy organization that supports high-growth, high-tech startups. Engine works with federal, state, and local government; international advocacy organizations; and a community of growth-oriented technology startups nationwide to support the development of technology entrepreneurship. Engine conducts research, organizes events, and spearheads campaigns to educate elected officials, entrepreneurs, and the general public on issues vital to fostering technological innovation, including improving patent quality. Engine works with many patent owners and innovators, including artificial intelligence startups. Engine has seen the detrimental impact improperly designed patents can have on innovation and appreciates the Patent Office's attention to this important subject area.

The Electronic Frontier Foundation – EFF is a non-profit civil liberties organization that has worked for more than 25 years to protect consumer interests, innovation, and free expression in the digital world. EFF and its more than 30,000 dues-paying members care deeply about ensuring that intellectual property law in this country serves the goal set forth in the Constitution: promoting the progress of science and technological innovation. To ensure the voices of consumers, end users, and developers are heard, EFF has often provided comments on behalf of the public's interest in the patent system to the USPTO, including on patent-eligibility requirements and their impact on innovation in the software industry.

Introduction

Artificial intelligence ("AI") is a quintessential disruptive technology. It has already significantly affected aspects of our everyday lives, from healthcare[1] to entertainment.[2] And it is difficult to imagine an industry or sector AI will not touch in the future. While many AI technologies are already ubiquitous, we are still in the early stages of an AI revolution, with myriad new advanced techniques and commercial applications on the horizon.

But, despite its transformative tendencies, AI does not need to disrupt the U.S. patent system. Patents have adapted to accommodate revolutionary technologies in the past, such as computer software and genetic engineering. While our patent policies should account for the value of emerging AI technologies – and we commend the Patent Office for seeking public input – the U.S. patent system does not now need substantial changes to accommodate AI.

Existing statutes, regulation, guidance, and case law map well onto the types of AI inventions commonly produced today and on the immediate horizon. The following comments focus on how the current frameworks for subject matter eligibility under § 101 and enabling disclosure under § 112 should apply to AI inventions for promoting progress and establishing high-quality patents in the field.

AI technologies perform tasks that conventionally require human intelligence, such as learning, reasoning, and perception. Usually, these technologies are implemented as computer software or hardware. AI is a broad discipline, including technologies such as expert systems, fuzzy logic, and

[1] Brian Kalis, Matt Collier & Richard Fu, *10 Promising AI Applications in Health Care,* Harvard Business Review (May 10, 2018), https://hbr.org/2018/05/10-promising-ai-applications-in-health-care.

[2] *Meson: Workflow Orchestration for Netflix Recommendations,* Netflix Technology Blog (May 31, 2016), https://medium.com/netflix-techblog/meson-workflow-orchestration-for-netflix-recommendations-fc932625c1d9.

robotics. But, regardless of the specific technology, most AI innovations today involve machine learning methods.

Machine learning methods solve problems without being explicitly programed. These methods have roots in statistical modeling and largely use statistical methods. Generally, both statistics and machine learning develop mathematical models from analyzing the inputs and outputs of a process. However, whereas traditional statistics tries to define a model of the process itself, machine learning methods try to predict the outputs of that process without trying to model or understand how it works.[3] By treating the process as unknown while trying to functionally approximate it, machine learning methods can learn to perform incredibly complicated and not well-understood tasks, such as object detection, that would be difficult, if not impossible, to explicitly program.

The basic development process of an AI system that uses machine learning is:

1. Defining the problem to solve.
2. Gathering and preparing data for training the AI system.
3. Selecting and building the machine learning model(s) for use in the AI system.
4. A loop of training, testing, and tuning the AI system until it is either ready for deployment or some aspect of it needs to be redesigned.

See the Appendix for diagrams of this development process, a general machine learning model, and the basic model training process.[4] Common

[3] Leo Breiman, *Statistical Modeling: The Two Cultures*, 16 Statistical Science 199 (2001), http://www2.math.uu.se/~thulin/mm/breiman.pdf.

[4] To create the diagrams, the commenters relied on two sources: Victor Roman, *How to Develop a Machine Learning Model from Scratch*, Towards Data Science (Dec. 23, 2016),

examples of machine learning methods include neural networks, support vector machines, and decision trees.

The request for comments refers to two types of AI inventions: Inventions produced by AI itself and inventions that use AI to solve tasks. AI-produced inventions may pose many difficult and interesting questions for the patent system, such as whether an AI can legally be the inventor, which entity – if any – should own the patent, and how an AI-inventor affects obviousness questions and the definition of an ordinarily skilled artisan.

From the perspective of promoting innovation, however, the second question – how the patent system should handle inventions that use AI – is currently more pertinent. Inventors file an increasing number of AI patent applications each year,[5] with claims we consider to fall into three general categories: (1) Methods for developing AI systems to perform specific tasks; (2) technical improvements to the development process; and (3) the AI system created at the end of the process. Depending on the scope of the claims and detail of the specification, inventions in each category may encounter issues with subject matter eligibility or enabling disclosure in a patent application.

These comments will address three points responsive to questions from the Patent Office's request for comments. First, regarding subject matter eligibility and application of the *Alice/Mayo* test, the Patent Office should prevent the issuance of patents which generically claim the performance of abstract ideas using generic, conventional AI technology, as such overbroad claiming would preempt downstream innovation (Question 5). Second, AI invention can be an unpredictable art, so satisfying the enablement requirement should require detailed descriptions of both the model

https://towardsdatascience.com/machine-learning-general-process-8f1b510bd8af; Ayush Pant, *Workflow of a Machine Learning Project*, Towards Data Science (Jan. 10, 2019), https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94.

[5] *WIPO Technology Trends 2019: Artificial Intelligence*, World Intellectual Property Organization (2019), https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf.

and the input (Question 7). Finally, any changes to the patent system for AI should promote innovation and disclosure (Question 11).

5. Are there any patent eligibility considerations unique to AI inventions?

The current framework for assessing patent eligible subject matter should apply to patents on AI

AI inventions do not require unique patent eligibility law or policy considerations. Recently decided cases, especially since *Alice*,[6] have helped to reduce the proliferation of weak, overbroad patents by refining subject matter eligibility standards. The Patent Office should be careful these standards do not regress when considering AI inventions. The *Alice/Mayo* test[7] should apply to AI inventions just as well as any other invention. While an AI invention embodied in software is more likely to be directed towards an abstract idea than an AI invention embodied in specific hardware, resolving this issue should require no special analytical steps outside applying the *Alice/Mayo* test as usual.

AI in Specific Hardware – Inventions where AI is implemented in specific hardware, rather than a generic computer, should not encounter novel eligibility issues under the *Alice/Mayo* test. For example, claims involving AI microprocessors or machines that use trained AI to complete a specific task may not be directed towards abstract ideas. Such claims would not recite any mathematical formula, method of organizing human activity, or mental processes. And, even if they did, such inventions often implement AI as part of a greater system in which the hardware provides a

[6] Alice Corp. Pty. Ltd. v. CLS Bank Int'l, 573 U.S. 208 (2014).

[7] *See Manual of Patent Examining Procedure, Section 2106,* U.S. Patent & Trademark Office (last revised Jan. 2018) (citing *Alice*, 573 U.S. at 216; Mayo Collaborative Servs. v. Prometheus Labs., Inc., 566 U.S. 66, 71 (2012)).

technological solution to a technological problem. We expect that for many hardware AI inventions, there would be few cases that progress past the first step of the *Alice/Mayo* test.

AI in Software – Most AI inventions are a system of algorithms implemented in computer software that run on generic computer hardware. In the context of subject matter eligibility, these inventions should be considered a subcategory of software inventions and subjected to standard software patent analysis. Valid AI patent claims, for example, could include methods for training an AI to perform a task or improvements to the training process itself.

AI software inventions should not be any more patentable than typical software. If anything, AI patent claims are more likely to be directed towards abstract ideas under *Alice* and subsequent case law than other software inventions. This is primarily for two reasons:

First, unlike typical software that is programed with explicit constructions, AI software often uses self-learning algorithms to achieve tasks. Essentially, these algorithms – many of which are now conventional, widely available "off-the-shelf" technology – optimize mathematical models for approximating an opaque phenomenon from its inputs and outputs. Claiming such a process, without some additional limitations, can be the sort of data manipulation and generation the Federal Circuit considered patent-ineligible in *Digitech*[8] and *SAP America*.[9] Additionally, a claim just for the trained model itself should be ineligible because it is a mathematical representation of the relationship between the input and output data.

[8] Digitech Image Techs., LLC v. Electronics for Imaging, Inc., 758 F.3d 1344, 1351 (Fed. Cir. 2014) ("Without additional limitations, a process that employs mathematical algorithms to manipulate existing information to generate additional information is not patent eligible.").

[9] SAP Am., Inc. v. InvestPic, LLC, 898 F.3d 1161, 1163 (Fed. Cir. 2018) (holding "mathematical calculations based on selected information and the presentation of the results of those calculations" not patent eligible).

Second, AI inventions, by definition, try to perform tasks traditionally requiring human intelligence. Broad claims for AI inventions may thus be directed towards concepts analogous to human mental work, from conscious processes, like making predictions, to implicit processes, like perception. AI technologies are powerful information processing systems that can perform a wide variety of tasks. The simple idea of using an AI system to do a typical human mental task, without more, is not the type of inventive contribution warranting patent protection. As *Alice* established with computers, merely using AI to carry out an abstract idea, such as mental steps, should not be enough to transform an abstract idea into patent eligible subject matter.

Overall, software patents should be the ceiling for AI software patents. AI inventions can be subjected to the same analysis and case law as other software. While, by nature, AI software tends to be a little more abstract, AI inventions pose no subject matter eligibility issues outside the domain of the current frameworks of the U.S. patent system.

7. How can patent applications for AI inventions best comply with the enablement requirement, particularly given the degree of unpredictability of certain AI systems?

AI inventions are an unpredictable art that require extra detail for an enabling disclosure

Patents on AI systems must properly disclose both the model and the input to meet the enablement requirement. AI systems thrive in an uncertain world. They identify imperceptible patterns and make unexplainable predictions from noisy, complex, and incomplete information. Much like their environment, AI systems themselves are volatile. Their self-learning algorithms require minimal human intervention and often use random processes to explore possible solutions. It can be difficult to tell whether a system learned a generalized solution to a given problem or a solution that only works with its training data. There is no guarantee running the same algorithm on the same dataset will even reach the same result. These characteristics make AI inventions a highly unpredictable art.

In the patent context, this unpredictability can pose major problems for § 112's enablement requirement, especially for patents that claim either an AI system for doing a specific task or a method for training such a system. In addition to being an inherently random and unpredictable process, designing and training AI systems involves significant discretion and creativity. Even a person having ordinary skill in the art may need to engage in a lengthy trial-and-error process to recreate the invention without appropriately detailed disclosures. The patent system already has a framework in place to prevent such undue experimentation.[10] As with other unpredictable arts, AI inventions require a more specific description for an enabling disclosure than more predictable arts.

At minimum, patents relating to AI systems performing a task must disclose in detail both the machine learning models used and how to train them. AI's strength is that it can learn to perform tasks without being explicitly programmed. Popular machine learning methods, such as deep learning, effectively create black boxes that make predictions about the world without explaining how they do it. Thus, an AI system's usefulness and functionality happens outside an inventor's direct control and often understanding. Essentially, an inventor's role in the process is setting up the proper environment for an AI system to learn its task. Enabling patents must teach the public how to set up this environment.

[10] *See Manual of Patent Examining Procedure, Section 2164.03,* U.S. Patent & Trademark Office (last revised Jan. 2018) ("The amount of guidance or direction needed to enable the invention is inversely related to the amount of knowledge in the state of the art as well as the predictability in the art.").

Proper model disclosure includes the model architecture and objective function

Disclosing a model's architecture and objective function is necessary for an ordinarily skilled artisan to recreate the invention without undue experimentation. Even machine learning models of the same type can vary drastically. For example, convolutional neural networks, which are well suited for tasks with visual data, have an arbitrary number and arrangement of layers, each with a set height, width, and depth.[11] Patents must describe any variations from prior art model architectures in detail. Otherwise, a skilled artisan will be forced to conduct many experiments to find the correct architecture from countless possible ones. For the deep neural networks widely used today, this process becomes more difficult the more layers there are in the network. More layers generally help networks perform more complex tasks.[12] However, deeper layers' functionalities are also increasingly unpredictable and difficult to visualize and understand intuitively.[13] AI inventions using deep learning should contain extra guidance on how to put the model in a position to learn successfully without substantial trial and error.

In addition to the model architecture, enabling disclosures must also include the model's objective function. Objective functions mathematically define the task a model is supposed to perform. They measure the level of "success" for any given set of parameters. Any changes to the objective function mean the model technically is being optimized for a different task.

[11] Andrej Karpathy, *CS231n Convolutional Neural Networks for Visual Recognition,* (last visited Nov. 8, 2019), https://cs231n.github.io/convolutional-networks.

[12] Jason Brownlee, *How to Configure the Number of Layers and Nodes in a Neural Network,* Machine Learning Mastery (July 27, 2018), https://machinelearningmastery.com/how-to-configure-the-number-of-layers-and-nodes-in-a-neural-network.

[13] Jason Yosinski et al., *Understanding Neural Networks Through Deep Visualization,* Cornell Univ. Computer Science Dept. (June 22, 2015), https://arxiv.org/abs/1506.06579.

Despite being a relatively small part of the entire training process, the objective function needs to be exact to ensure the AI system is learning as intended. There are many well-documented and commonly used objective functions.[14] A model using one of those objective functions would only have to specify which one to enable a skilled artisan to recreate the AI system. However, any custom objective functions should either be expressed mathematically or in pseudo-code in the specification.

Proper input disclosure requires either a detailed description of a training data's structure or disclosing the training data itself

Describing the model, while necessary, is alone not sufficient for enablement. Proper patent disclosures must also include how an inventor trained the model, most significantly the model's input data. The input is so important to AI systems that phrases such as "garbage in, garbage out" and "data is the new oil" became mantras in the community. Managing the input is at least as crucial to an AI system's success and requires as much inventor ingenuity as preparing the algorithm.[15] Training an AI system, no matter how carefully designed it is, could be impossible without an appropriate dataset.

A patent that only discloses what sort of generic data is needed to train an AI system is not enough to meet the enablement requirement. Given the unpredictability of AI systems, a skilled artisan needs more guidance to avoid undue experimentation in recreating the proper input. Any of the

[14] Lars Hulstaert, *Understanding Objective Functions in Neural Networks* (Nov. 4, 2017), https://towardsdatascience.com/understanding-objective-functions-in-neural-networks-d217cb068138.

[15] *In Machine Learning, What is Better: More Data or Better Algorithms*, KD Nuggets (June 2015), https://www.kdnuggets.com/2015/06/machine-learning-more-data-better-algorithms.html; Chen Sun et al., *Revisiting Unreasonable Effectiveness of Data in Deep Learning Era*, Cornell Univ. Computer Science Dept. (July 10, 2017), https://arxiv.org/abs/1707.02968.

steps taken to prepare the data and the characteristics of the dataset's structure could prove dispositive in the system's ability to learn. This is especially important for tasks where the input itself is not readily understandable for humans, such as data with thousands of inputs or that is dimensionally reduced. An inventor does not necessarily know which aspects of any input are essential for the self-learning algorithms to optimize for their tasks. To meet the enablement requirement, a patent for an AI invention should disclose in detail what sort of data the model needed to learn.

An inventor has at least two options for sufficiently disclosing the data necessary for training an AI system. First, an inventor could describe the training data in the specification. The description should include all data characteristics necessary for the AI to learn the task. For example, for a dataset of facial images, these characteristics might include the quantity and demographics of the people, the lighting conditions and camera angles of the images, or any adjustment filters and sampling applied to the final input. If appropriate, the figures might also include some drawings of mock inputs. Such guidance, in contrast to a terse direction such as "collect a large sample of facial data," would allow a skilled artisan to recreate a similar dataset without first trying to figure out the data's necessary features through guessing and experimenting.

Alternatively, an inventor could opt to disclose the dataset directly. For example, if an AI system were trained using a commonly available dataset, such as ImageNet,[16] citing it would be a sufficient disclosure. An inventor might also self-publish the data. These data disclosure methods echo the patent procedure for biological deposits. Like AI systems, those self-replicating organisms function in invention-relevant ways outside the inventor's direct control. And, as with biological deposits, a dataset description may not convey all the necessary information to enable a skilled artisan to make or use an AI invention. The Patent Office could even create and maintain a

[16] ImageNet Homepage (last visited Nov. 8, 2019), http://image-net.org/index.

data depository. The public could request samples from the depository and inventors would be able to meet the enablement requirement without self-publishing the data.

Overall, the patent system already has the tools necessary to assess whether patent applications comply with the enablement requirement. AI inventions are a particularly unpredictable art, highly dependent on the inventor's design choices and the input data. Patent applications could greatly help enablement analysis by indicating how unpredictable the AI system and training process are. Assuming it met the eligibility requirement, a system implementing well-known models, like AlexNet,[17] on openly available datasets would require little extra disclosure because these systems are so well-documented. However, if a completely custom algorithm was trained under specific circumstances, even a skilled artisan would need to know how exactly to set up the model's environment to recreate the invention.

Because the enabling requirement is a high bar for AI inventions, a patent application with sufficient details for an enabling disclosure may signal that other requirements for patent eligibility have also been met. By requiring more detailed descriptions, it seems likely an AI patent application that meets the enablement requirement can also meet the written description requirement. Moreover, detailed descriptions in the specification must also correlate to precision in the claims. Precise claims are essential to make the public aware of what exactly the invention is, so they know if they infringe the patent. The enablement and written description requirements must be carefully and thoroughly vetted during prosecution to ensure that AI patents serve their notice and disclosure functions.

[17] Alex Krizhevsky et al., *ImageNet Classification with Deep Convolutional Neural Networks*, Neural Information Processing Systems Conference (2012), https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

11. Are there any other issues pertinent to patenting AI inventions that we should examine?

The AI field already exhibits substantial innovation and disclosure under the current patent framework

The U.S. patent system exists to promote innovation and disclosure. Over the past decade, few fields have experienced a greater explosion in innovation than AI. There are new developments and advancements in AI technology all the time, as investment, research, and interest in the industry increased drastically over the past few years.[18] AI development thrives everywhere, from universities and research institutions to startups and large companies. To remain globally competitive, the U.S. needs to keep investing in AI research and education.[19]

Notably, this innovation and growth in AI all happened under the current patent framework. AI development does not need extra incentive or stimulation. Thus, when considering any changes to how it evaluates AI patent applications, the Patent Office should carefully consider how those changes might impact this "Golden Age" of AI and whether they will hamper innovation. Easing the requirements for patentable AI inventions could have negative effects. In particular, the risk of permitting weak, overbroad patents, like those that plagued the patent system prior to *Alice*, could end up restricting downstream innovation in this field.

Additionally, the Patent Office should consider how any changes to the current patent framework might impact the high level of disclosure already

[18] AI Index Steering Committee, *The AI Index 2018 Annual Report*, Human-Centered AI Initiative, Stanford University (2018), http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf

[19] Tom Simonite, *China is Catching Up to the US in AI Research – Fast*, Wired (Mar. 13, 2019), https://www.wired.com/story/china-catching-up-us-in-ai-research.

in the AI development community. AI systems and tools are often distributed for free online under permissive licenses, whether they were developed by for-profit corporations, non-profit reach institutions, or a dedicated hobbyist. For example, TensorFlow,[20] one of the most powerful and widely used libraries for machine learning in the world, is distributed under the free and open source Apache 2.0 license,[21] which includes an explicit patent grant. Patents naturally support the community's open culture by promoting the disclosure of complex AI inventions that might otherwise remain secret in exchange for granting a temporary monopoly on the invention. However, patents without a proper enabling disclosure allow this monopoly without teaching the public how to make or use the invention. Granting these types of patents not only results in patents that fail under the statute, but also undercuts the open ideals of the community, whose work many in the field depend on and profit from.

Finally, the Patent Office should also understand the greater policy contexts of AI inventions when examining issues around AI patents. Opaque AI systems trained with biased datasets are negatively impacting the lives of already marginalized communities.[22] While addressing these issues is not necessarily within the Patent Office's mandate, the patent

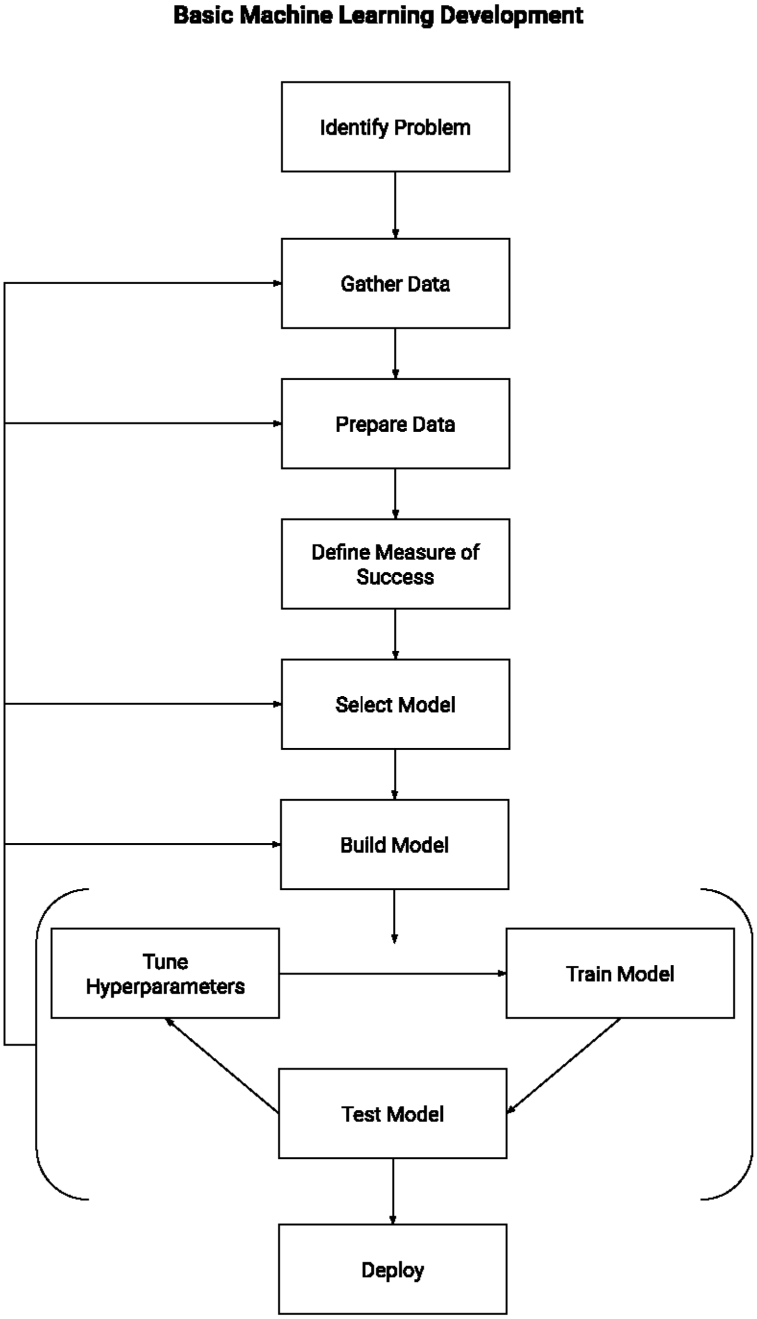
[20] TensorFlow homepage (last visited Nov. 8, 2019), https://www.tensorflow.org.

[21] *Apache License, Version 2.0*, Apache Software Foundation (last visited Nov. 8, 2019), https://www.apache.org/licenses/LICENSE-2.0.

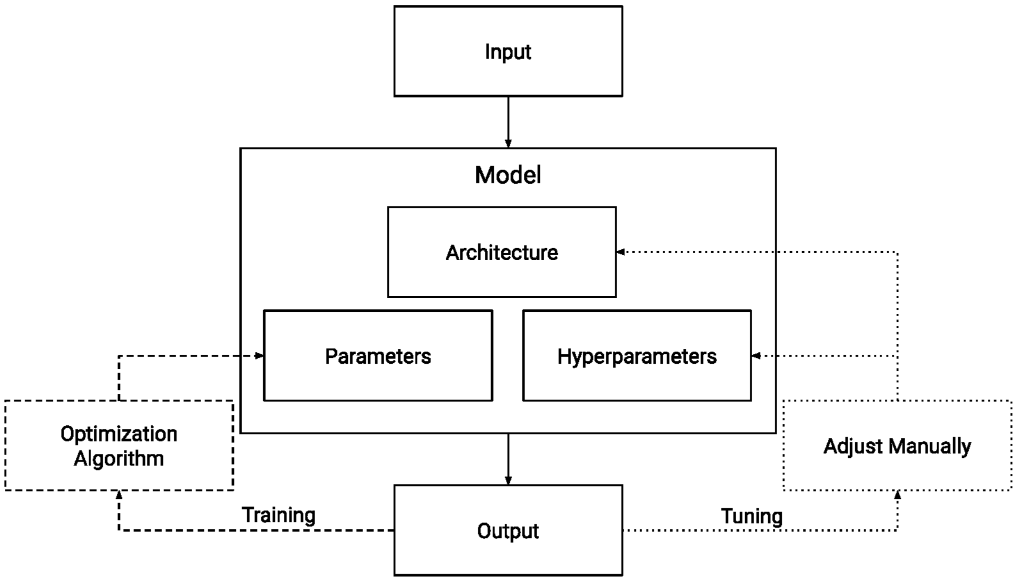
[22] *See, e.g.*, Julia Angwin et al., *Machine Bias*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Hal Hodson, *Police Mass Face Recognition in the US Will Net Innocent People*, NewScientist (Oct. 20, 2016), https://www.newscientist.com/article/2109887-police-mass-face-recognition-in-the-us-will-net-innocent-people; Natasha Singer, *Amazon Is Pushing Facial Technology That a Study Says Could Be Biased*, N.Y. Times (Jan. 24, 2019), https://www.nytimes.com/2019/01/24/technology/amazon-facial-technology-study.html; Jeffery Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*, Reuters (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G; Aaron Rieke & Corrine Yu, *Discrimination's Digital Frontier*, Atlantic (Apr. 15, 2019), https://www.theatlantic.com/ideas/archive/2019/04/facebook-targeted-marketing-perpetuates-discrimination/587059.

system itself can nonetheless have an impact. Strong patents with precise claims can better help the public understand how AI systems make decisions that impact their lives. And properly disclosing the data used to train AI systems in order to meet the enablement requirement can give the public the opportunity to check for bias in the dataset or data collection process. With AI inventions, patents have a special opportunity to promote not only the progress of science and useful arts, but also the public's general welfare.

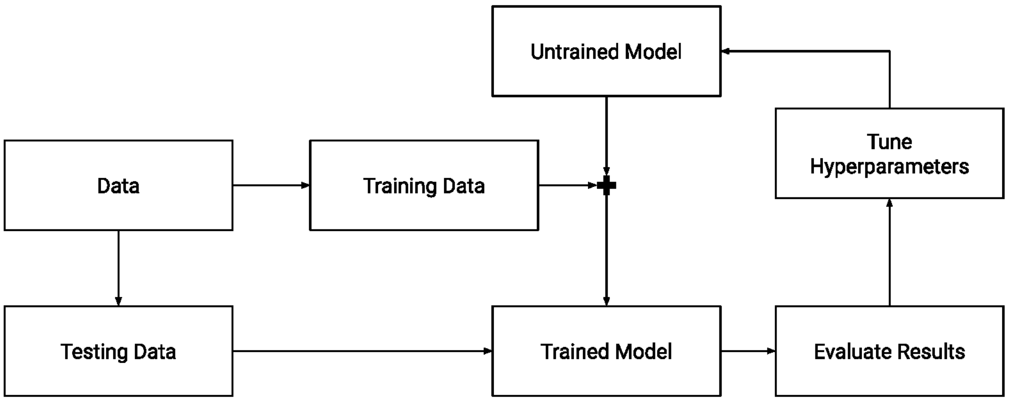
Appendix


Basic Machine Learning Development
Identify Problem
Gather Data
Prepare Data
Define Measure of Success
Select Model
Build Model
Tune Hyperparameters
Train Model
Test Model
Deploy

Generalized Machine Learning Model



Basic Model Training Process



Appendix B

Engine

January 10, 2020

Andrei Iancu
Under Secretary of Commerce for Intellectual Property
Director of the United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450

VIA EMAIL

Re: Comments of Engine Advocacy in Response to *Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation*, Docket No. PTO-C-2019-0038

Dear Director Iancu:

Engine is a non-profit technology policy, research, and advocacy organization that bridges the gap between policymakers and startups. Engine works with government and a community of thousands of high-technology, growth-oriented startups across the nation to support the development of technology entrepreneurship. We appreciate the United States Patent and Trademark Office's ("USPTO") attention to pursuing policies that foster innovation. And we further appreciate the opportunity to submit these comments regarding the impact of artificial intelligence ("AI") technologies on intellectual property ("IP") law and policy, as well as the impact IP law and policy could have on emerging AI technology.

Existing statutory language and case law appropriately permit the use of potentially copyrightable material in training, tuning, and testing AI systems. These comments address the third question raised in the USPTO's request, and explain how the law applies to render such use lawful.

Overall, the relevant legal and policy frameworks are working well. To the extent any changes are pursued, those changes should only codify the legal standards as articulated below and instill increased certainty that these standards will govern the development of AI technology. But if future developments in AI render the current law unduly ambiguous or unclear about the infringement liability associated with AI processing content, then it may be appropriate to amend relevant statutes to confirm that such uses are noninfringing.

I. **Introduction**

Under existing law, it is lawful to ingest content and use it to train, tune, and/or test AI systems. Such uses either involve content that is ineligible for copyright protection, involve uses that are not infringing, or involve fair uses of content.[1] While there may be ways to streamline or codify the law to make it more certain and predictable, existing law is working well.

The USPTO has asked how copyright law does (or should) apply to the process of AI ingesting content and using it as the AI system learns its functions. Ingesting involves bringing data into an AI system, and, for example, filtering, transforming, integrating, and/or validating it as necessary to ensure data quality.[2] And in prior comments, we used the terms training, tuning, and testing to describe the processes through which an AI system learns its functions, and use that same language in these comments.[3] While each disparate approach to and application of AI might call for slightly different treatment under copyright law, there are common features which make ingesting content and using it to train, tune, or test lawful.

AI is already pervasive in our everyday lives, and there are ever (and rapidly) expanding advanced AI technologies and commercial applications on the horizon. AI systems are designed, and will continue to be designed, to solve many different types of problems in reliance on many different types of data. Similarly, content that is potentially eligible for copyright protection[4] is ubiquitous and varied. Copyright protection can apply to *any* "original works of authorship fixed in [a] tangible medium of expression."[5] While underlying facts, data, ideas, etc., are not eligible for copyright protection, the broader works that contain underlying facts might have expressive elements that render such works copyrightable.[6] As such, the data used to train, test, and tune AI systems may be taken from content that is potentially eligible for copyright protection.

An example is useful for understanding why ingesting copyrighted material should be lawful, and why changing the law would be problematic. Internet platforms and service providers are

[1] One possible exception, discussed in Part III below, turns on whether the USPTO considers "ingesting" to include gathering content (i.e., collecting the material that is compiled into the dataset). If data gathering, or collecting, is part of ingesting, and if an AI system's developer creates unauthorized copies of copyrighted material at that point, that could constitute infringement. The infringement analysis would still depend, at least, on whether the gathered content is in fact eligible for copyright protection and whether the system's use was a fair use.

[2] *See, e.g.*, Alistair Croll, *The Feedback Economy*, *in* PLANNING FOR BIG DATA 1, 4 (Edd Dumbill ed., 2012) (defining ingesting and cleaning); SUNILA GOLLAPUDI, PRACTICAL MACHINE LEARNING 70 (describing the data ingestion layer).

[3] *Comments of Engine Advocacy & The Electronic Frontier Foundation*, PTO-C-2019-0029, at 4 (Nov. 8, 2019).

[4] For the purposes of these comments, the term *content* is used as shorthand for *content that is potentially eligible for copyright protection*.

[5] 17 U.S.C. § 102(a).

[6] *See* 17 U.S.C. § 102(b).

under increasing pressure to affirmatively identify user-generated copyright infringement. While current technology is (putting it mildly) far from perfect, many companies are trying to develop AI systems to improve automatic detection of potential infringement (or at least narrow the set of potentially infringing material that requires human review).[7] To develop and train those systems, companies need to feed copyrighted material into algorithms—a system that needs to draw lines between infringement and noninfringement can only do that if it knows what infringement looks like. But if using copyrighted material to train, tune, and test such a system is itself infringement, it destroys the incentive (or at least business case) for doing the development work, because the development itself is costly copyright infringement.

Policymakers should not change the status quo. If ingesting content were unlawful, that would open AI startups and innovators to unbearable costs and risks. It would ultimately slow, and could even stall, domestic AI development and along with it American leadership in the field. Overall, while there may be areas of relevant copyright law and policy that are not entirely settled, the existing interpretations described herein are working well and should continue to control.

II. **For datasets that exclude any expressive content, all use is lawful**

All AI systems involve volumes of data at the foundation of their development. And in many cases, that data may be taken from copyrighted works. But if the data is just that—data—and not anything expressive, the entire copyright question is moot because there is no copyrighted material that could even be infringed. For example, a facial recognition system may rely on a dataset of tightly cropped images of faces extracted from photographs. If the expressive contents/portions of the photographs are removed when the dataset is created, the data may not even be eligible for copyright protection.[8] For AI systems that exclusively analyze factual information extracted from content, the copyright inquiry should end there, because that factual information is not copyrightable.

[7] *See, e.g.*, Adam Satariano, *Europe Adopts Tough New Online Copyright Rules Over Tech Industry Protests*, N.Y. TIMES, Mar. 26, 2019; Krista L. Cox, *Automated Copyright Filtering Removes Public Domain Mueller Report From Platform*, ABOVE THE LAW (May 9, 2019, 11:17 AM), https://abovethelaw.com/2019/05/automated-copyright-filtering-removes-public-domain-mueller-report-from-platform/; Evan Engstrom & Nick Feamster, *The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools* (Mar. 2017), https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/58d058712994ca536bbfa47a/1490049138881/FilteringPaperWebsite.pdf.

[8] *See, e.g.*, Daryl Lim, *AI & IP: Innovation & Creativity in an Age of Accelerated Change*, 52 AKRON L. REV. 813, 850-51 (2018) (noting that facial recognition databases compiled from news images may not even invoke fair use if the portions of the photos taken is minimal); Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 67-68 (2017) (similar).

III. Ingesting data and using it during training, tuning, and testing is not infringing

AI systems are trained, tuned, and tested using volumes of content, and that use is not infringement. That sort of use is most analogous to a person reading a book, listening to a song, or viewing a painting. None of those personal uses are copyright infringement, and AI ingesting and processing content should be treated similarly.

Copyright law has traditionally "left reading, listening, and viewing unconstrained."[9] Not only can people can read books, listen to records, and look at art in museums without running afoul of the law. They can jot down their impressions while they read or recall a movie quote without violating an author's rights.[10] The law excludes myriad similar, technology-based uses from copyright infringement liability. For example, people are allowed to copy digital music files from their computer to an MP3 player.[11] And many of us back-up our hard drives on a periodic basis, making archival copies of any copyrighted material we own. The law and society (either explicitly or effectively) treat these as lawful.[12]

As we have previously noted, "AI technologies perform tasks that conventionally require human intelligence, such as learning, reasoning, and perception."[13] AI technology is performing human-like tasks, and it is engaging with content like humans do when they read, listen, or view. Therefore, when an AI system analyzes content during training, tuning, or testing, just like a lawful personal use, that should be outside the scope of infringement.

The basic development process for an AI system includes ingesting and preparing content as well as a loops of training, testing, and tuning when the content is analyzed and processed.[14] Those uses of content do not involve reproduction, distribution, performance, public display, or creation of derivative works within the meaning of the statute.[15] Admittedly, during ingestion, training, testing, and tuning content "may be copied, emulated, and re-copied thousands of times during

[9] Jessica Litman, *Lawful Personal Use*, 85 TEX. L. REV. 1871, 1882 (2007).

[10] *See, e.g.*, *id.* at 1893 ("It would plainly be unconstitutional to prohibit a person from singing a copyrighted song in the shower or jotting down a copyrighted poem he hears on the radio.") (quoting Justice John Paul Stevens).

[11] Recording Indus. Ass'n of Am. v. Diamond Multimedia Sys., Inc., 180 F.3d 1072, 1079 (9th Cir. 1999) ("the purpose of the Act is to ensure the right of consumers to make analog or digital audio recordings of copyrighted music for their private, noncommercial use") (citations omitted).

[12] *See, e.g.*, Litman, *supra* note 9, at 1895-1898, 1902 (discussing statutory exceptions to copyright infringement, such as making backup copies of computer programs (section 117) and making noncommercial copies of recorded music (section 1008), exceptions carved out in case law, and personal uses that do not fall within an explicit exclusion but nonetheless constitute common personal uses of content that would, uncontroversially, be considered noninfringing).

[13] *Comments of Engine Advocacy & The Electronic Frontier Foundation*, PTO-C-2019-0029, at 3 (Nov. 8, 2019).

[14] *Id.* at 4, 17-18.

[15] *See* 17 U.S.C. § 106 (defining a copyright owner's exclusive rights).

the learning process."[16] This could include creating ephemeral copies, that are so transitory in nature that they do not even constitute creating a *copy* as defined in the statute.[17] And depending on the AI method or model used, humans may have no insight into how the AI system is analyzing and processing data, nor what the intermediary content looks like.[18] For these hyper-technical instances of copying or modification, "the spirit of the copyright statute seems to exempt this type of copying."[19] Just like courts "refuse to entertain discovery with respect to early drafts of a noninfringing final work" on the basis that those interim and unpublished drafts are not infringements,[20] the interim status of content during the AI training process should be free from infringement scrutiny.

The one caveat to this assessment of noninfringing uses turns on the definition of "ingesting." If the USPTO understands "ingesting" content to encompass the collection of content in the first instance for inclusion in a dataset, it is possible that developers would make digital copies of content during that collection (or data gathering) process. For some AI applications, and for some developers (particularly those that do not have in-house access to data), it may be necessary to copy content "in order to process [it] as grist for the mill, raw materials that feed [] algorithms."[21] Because, at the very least, that initial copying would qualify as fair use, it is therefore also lawful.[22]

IV. <u>Even if such uses were potentially infringing, they would be lawful fair uses</u>

In the interest of promoting progress and innovation, it would be better to resolve that AI use of content is lawful because it is a noninfringing use. The alternative, fair use, is decided on a case-by-case basis.[23] Proceeding through litigation to establish that a specific use is fair is costly and not dispositive for all future (even similar) uses of data.[24] So while it is a fair use for AI to ingest and process data, it is more efficient to conclude that such uses are not even infringing.[25]

When content is used to train, tune, and/or test an AI system, if ingesting and processing that content were determined to be an infringing use then it would a lawful fair use. Numerous courts have applied the fair use factors to similar technology, and have consistently found those to be fair uses. The application of each fair use factor will vary, depending on what problem the AI

[16] Sobel, *supra* note 8, at 62-63.
[17] Sobel, *supra* note 8, at 62-63 (citing cases); *see also* 17 U.S.C. § 101 (defining "copies").
[18] *See, e.g.*, Davide Castelvecchi, *Can We Open the Black Box of AI?*, 538 NATURE 21 (2016).
[19] Sobel, *supra* note 8, at 63.
[20] Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. 1607, 1635-36 (2009).
[21] *Id.* at 1608.
[22] *Infra* part IV.
[23] *E.g.*, Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 577 (1994).
[24] *See, e.g.*, Litman, *supra* note 9, at 1902-03.
[25] *Supra* part III.

system is designed to solve, what content it uses, and how that content is gathered and prepared. But, as described below, themes readily emerge and the outcome is consistent.

Factor (1): Purpose and character of the use. With this first factor, courts must determine if use of content "merely supersede[s] the object of the originals or instead add[s] a further purpose or different character."[26] And in the case of AI, it does not supersede the original content it relies on in training, tuning, and testing. Instead, AI's use of the content adds a further purpose or different character.

While "[t]ransformative use is most obvious when [a] work is itself transformed[,] in many cases courts have held that the mere recontextualization of a copyrighted work from one expressive context to another is sufficient to sustain a finding of fair use."[27] For example, making a digital copy of a book so that it is easier for people to search within books is transformative.[28] Making thumbnail copies of images to help index and improve access to content on the Internet is also transformative.[29] And archiving student-written term papers within a database for an online plagiarism detection technology is transformative.[30] AI likewise recharacterizes content.

Courts also look at the commercial nature of a use when assessing this first factor. But a commercial motivation cannot outweigh an otherwise transformative use,[31] so AI developed with a commercial application in mind can still be a fair use. Like the use of thumbnail images in a search engine (which is both commercial and transformative), AI systems do not use the content in datasets to directly profit, and are not making a profit off of those individual pieces of content. Instead, each piece of content in an AI dataset is among thousands (or many more) elements being used in a commercial endeavor.[32] So even where an AI developer has an overarching commercial purpose, that commercial aspect does not defeat its transformative nature.

Factor (2): Nature of the copyrighted work. The nature of the content used will vary for each AI system. However, this factor rarely plays a significant role—standing alone—in determining fair use.[33] Especially because the use of content to train, tune, and test AI systems is so transformative, even if the content is highly creative, this factor should not tip the scales against a fair use finding.[34]

[26] Kelly v. Arriba Soft Corp., 336 F.3d 811, 818 (9th Cir. 2003).
[27] Sag, *supra* note 20, at 1646.
[28] Authors Guild v. Google, Inc., 804 F.3d 202, 216-17 (2d Cir. 2015).
[29] *Kelly*, 336 F.3d at 818.
[30] A.V. *ex rel.* Vanderhye v. iParadigms, LLC, 562 F.3d 630, 640 (4th Cir. 2009).
[31] *Authors Guild*, 804 F.3d at 219.
[32] *E.g.*, *Kelly*, 336 F.3d at 818.
[33] *E.g.*, *Authors Guild*, 804 F.3d at 220 (citing WILLIAM F. PATRY, PATRY ON FAIR USE § 4.1 (2015)).
[34] *E.g.*, *Authors Guild*, 804 F.3d at 220.

Factor (3): Amount and substantiality of portion used. Here again, the amount of content each AI system uses will vary, but even if an AI system uses entire copyrighted works during the training, tuning, or testing processes, it can still qualify for fair use.[35] What matters is "the amount and substantiality of what is [] made accessible to a public for which it may serve as a competing substitute."[36] And in the context of AI ingesting or processing content, the answer is *none*. When an AI system is using content during training, testing, or tuning, it does not make anything accessible to the public, certainly no competing substitutes. This third factor also weighs in favor of finding fair use.[37]

Factor (4): Effect upon the potential market. It is unlikely that AI ingesting data will harm the market for the original work, because copyrighted works are developed to entertain, impress, or inform human audiences. And an AI system's use of a piece of content as part of the training, testing, or tuning process would not harm the creator's ability to sell or license the original content. An AI system does not sell, license, or even make publicly available the underlying original content.[38] In any event, because the purpose and character of AI's use of content is highly transformative, that highly transformative nature suggests there will not be market harm.[39]

V. <u>Allowing AI systems to ingest content and be trained free from copyright liability promotes innovation</u>

Existing law, as described above, is working well. And there are strong policy reasons against making changes. As one scholar noted, "[h]ow copyright law treats the use of [AI training] datasets will determine whether AI-generated works can reliably develop without a constant threat of litigation."[40] Changing copyright law or policy in a way that creates liability when AI systems ingest content would put innovation at risk.[41]

[35] *Authors Guild*, 804 F.3d at 221 ("Complete unchanged copying has repeatedly been found justified as fair use when the copying was reasonably appropriate to achieve the copier's transformative purpose and was done in such a manner that it did not offer a competing substitute for the original.")

[36] *Id.* at 222.

[37] *See, e.g.*, *id.* at 221-22.

[38] *See, e.g.*, Kelly v. Arriba Soft Corp., 336 F.3d 811, 818 (9th Cir. 2003) (finding use of thumbnail images in search engine did not harm photographer's ability to sell or license full-sized images); A.V. *ex rel.* Vanderhye v. iParadigms, LLC, 562 F.3d 630, 643 (4th Cir. 2009) (finding that archive of student-written papers in an online plagiarism detection tool "did not serve as a market substitute or even harm the market value of the works").

[39] *See, e.g.*, *Kelly*, 336 F.3d at 818 (citing *Campbell*, 510 U.S. at 586-87).

[40] Lim, *supra* note 8, at 847-48.

[41] *See, e.g.*, Christian Handke et al., *Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research*, in New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust 120 (B. Schmidt & M. Dobreva eds., 2015) (finding that data mining-related research is increasing globally, but for countries in which data mining requires consent from copyright owners, data mining reflects a smaller share of academic output).

Developing AI is a national priority, which means promoting innovation and not creating new hurdles.[42] Everyone who generates any content arguably has some claim to copyright protection (and such content can range from highly expressive paintings to largely-factual academic papers that are at least expressive in part). It would be untenable to require innovators developing new AI technology to assess the copyright status of every piece of content and every data source feeding into an AI system. If developers then had to obtain licenses to any content or data sources, it would magnify the problem enormously.[43] If developers faced those sorts of burdens when compiling datasets, it would slow (and could stall) progress.

Relatedly, the alternative to licensing content is using it without permission and risking infringement litigation. The cost of a single copyright infringement suit, where statutory damages of $150,000 are available as a matter of course, could be ruinous for a startup.[44] But "[b]ecause machine learning datasets can contain hundreds of thousands or millions of works, an award of statutory damages could cripple even a powerful company."[45] These costs and risks would scare many innovators, companies, and investors away from developing new AI systems.[46]

Lawful use of content can also help prevent bias problems in AI. If it is difficult for developers to navigate the copyright landscape when developing datasets, it could result in them leaving certain content out and creating a non-comprehensive dataset subject to bias. Access to content free from claims of copyright infringement gives developers access to more diverse, less biased content. For example, recently created (and therefore currently copyrighted) works may be less susceptible to inherent or traditional biases, when compared to public domain content that was created when the U.S. was less racially diverse and had more gender disparity.[47] And "[b]ias in AI may be exacerbated by a restrictive fair use doctrine," because if training data "is protected by copyright, those who use [it] do so secretly, preventing biases from being uncovered."[48]

[42] *See, e.g.*, *Artificial Intelligence for the American People*, WHITEHOUSE.GOV, https://www.whitehouse.gov/ai/ (last visited Jan. 7, 2020) (recognizing that "America has long been the global leader in this new era of AI" and describing five pillars for advancing AI in the U.S., including "removing barriers to AI innovation"); National Security Commission on Artificial Intelligence, *Interim Report*, at 1 (Nov. 2019), *available at* https://drive.google.com/file/d/153OrxnuGEjsUvlxWsFYauslwNeCEkvUb/view (describing AI as "integral to the technological revolution that we are now experiencing," and expressing concern that "America's role as the world's leading innovator is threatened").

[43] Sag, *supra* note 20, at 1659-61 (describing high transaction costs that would be encountered if Internet search engines had to obtain copyright clearance for all searchable content).

[44] 17 U.S.C. § 504(c).

[45] Sobel, *supra* note 8, at 80; *see also* Lim, *supra* note 8, at 847-48 (similar).

[46] *See, e.g.*, Sobel, *supra* note 8, at 80-81.

[47] *See, e.g.*, Louise Matsakis, *Copyright Law Makes Artificial Intelligence Bias Worse*, VICE (Oct. 31, 2017, 12:00 PM), https://www.vice.com/en_us/article/59ydmx/copyright-law-artificial-intelligence-bias (citing Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018)).

[48] Lim, *supra* note 8, at 854.

Moreover, bias problems could arise if copyright owners were in a position to dictate when and who can develop AI technology and for which purposes.

Finally, increasing the burden on innovative companies that need access to content and data to develop AI technology disproportionately affects startups and entrenches big incumbents. Big technology companies have many users and troves of in-house data they can use to develop new AI systems. Because they own the necessary content, they would not have to worry about infringement.[49] These large companies also have significant bargaining and purchasing power for acquiring large volumes of content. Startups, on the other hand, who often must look externally for data sources, have to pull-in data from content that might be subject to copyright claims. They need to be able to do this without fear of infringement accusations.

AI development is thriving, and the recent explosion in promising AI technology occurred under the current copyright framework. Before making any changes to this framework, policymakers should carefully consider how any changes would impact the current trajectory of ubiquitous, varied, and exciting AI technologies. And policymakers should avoid making any changes that might hamper innovation.[50]

[49] *See, e.g.*, Steve Lohr, *At Tech's Leading Edge, Worry About a Concentration of Power,* N.Y. TIMES, Sept. 26, 2019.

[50] *See, e.g.*, *Comments of Engine Advocacy & The Electronic Frontier Foundation*, PTO-C-2019-0029, at 14-15 (Nov. 8, 2019) (addressing how changes to current patent framework might restrict AI innovation).