Generative AI, Ingenuity, and Law

Joseph R. Carvalko Jr., *Member, IEEE*

***Abstract*—This paper discusses generative pre-trained transformer technology and its intersection with forms of creativity and law. It highlights the potential of generative AI to change considerable elements of society, including modes of creative endeavors, problem-solving, employment, education, justice, medicine, and governance. The author emphasizes the need for policymakers and experts to join in regulating against the potential risks and implications of this technology. The European Commission has taken steps to address the risks of AI through the European AI Act (EIA), which categorizes AI uses based on their potential harm. The legislation aims to ensure scrutiny and control in extreme cases like autonomous weapons or medical devices. However, the author criticizes the lack of meaningful AI oversight in the United States and argues that time has come for government to step in and offer meaningful regulation given the technology's (1) rate of diffusion (2) virtually uncountable product permutations, the purposes, extent and depths to which it is anticipated to penetrate institutional and daily life.**

***Index Terms*—Artificial intelligence, computation and language, deep learning neural networks, NLP, LLM, OpenAI, ChatGPT, generative AI, generative pretrained transformer, transformer-based AI, European AI Act, EIA, technology ethics.**

> "Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks." Stephen Hawking

I. Introduction

IF QUANTUM mechanics, computers, and genetic engineering were the paradigm shifts that defined the twentieth century, artificial intelligence will undoubtedly define the twenty-first. For over a year I have watched handwringing by everyone from my social media contacts, to TV talking heads and U.S. senators concerned by the latest deep learning neural networks referred to as generative pre-trained transformers, or generative AI. The technology, developed by among others, Microsoft, Google, Meta, OpenAI, and Anthropic, takes form in what are denominated large language models, or LLMs, that use natural language text input to create content having different focuses and capabilities.

As generative AI technology becomes accessible to billions of people throughout the world it will increase the potential AI has to dramatically change the way a society functions. As observed with other world-changing technologies, such as fossil fuels and the Internet, unless we move quickly toward regulation, the opportunity window may close and thereafter it will be nigh impossible to course correct. In this regard, I share the clarion call by ethicists, technologist and policymakers that we ascertain what if any harmful consequences flow from the new technology and if so, put the topic of responsible innovation and balanced regulation front and center of our public discourse.

Part of appreciating the urgency for regulation, requires that we understand the broad outlines of the technology, i.e., what it does, how it works, and determine where and how it may compromise fundamental human rights or ethical institutional practice. While generative AI offers an instrument for good, one that promises to contribute positively to progress and well-being, it has considerable potency to cause economic disruption and a diminution in personal potential in myriad ways. It is essential that we recognize the impact of inaction. But it is also essential that we choose wisely the extent to which we act to limit the reach of a technology that has all the earmarks of transforming the lives of citizens and revolutionizing the practice of science, engineering, business, education, medicine, and law. As has become apparent in these early days of the technology's exploitation, generative AI exhibits all the signs of a double effect, the theory that "a course of action might have a variety of ethical effects, some 'good' and some 'bad.' It can be seen as a way of balancing consequentialist and deontological approaches to ethics$\cdots$ best explained through the classic thought experiment: the Trolley problem [1]." By analogy, generative AI technology increases the spectrum of creative output, while devaluing the human contribution in a wide range of artistic, musical and literary constructions. Unless controlled, it may have the effect of shifting much of the intellectual and creative kernel of human capital to autonomous agencies. Let us delve deeper into what we do and do not know about this new development, so that we can better assess where we need to put our energy and resources to abate the unintended and inadvertent side effects that it may cause.

The salient feature of a generative AI LLM is that it responds to simple language prompts by producing various content, such as emails, essays, poetry, fictional stories, images, or computer code for executing novel apps. Products such as OpenAI's ChatGPT, Google's Gemini, Meta Platforms' Llama 3, Microsoft's Copilot and Phi-3, and Anthropic's Claude are but a few of the latest offerings, each designed for redrafting and summarizing documents, generating functional code, and in some instances more specifically to analyze medical images, diagnose medical conditions or assist in contract drafting. Other products such as Midjourney, MuseNet, and Lyrical Labs attract users having interests in

Manuscript received 14 July 2023; revised 25 March 2024 and 6 June 2024; accepted 9 June 2024.

The author is with the Technology and Ethics Research Working Group, Interdisciplinary Center for Bioethics, Institution for Social and Policy Studies, Yale University, New Haven, CT 06520 USA (e-mail: joseph.carvalko@yale.edu).

Digital Object Identifier 10.1109/TTS.2024.3413591

artwork and graphics or songwriting, lyrics generation, and even music composition, respectively. On the surface, none of these activities would appear disruptive to modern life. But in these few descriptions of what these products offer, we are inevitably drawn to issues that implicate the future for artists, writers, and musicians, such as the manner by which they create or acquire ownership in the fruits of its output or the intellectual property utilized in training the generative AI model itself. In other instances, the technology is poised to restructure the delivery of medical services or the practice of law. But the potential for disruption does not stop here.

The current landscape involving generative AI can be systematically classified into various sectors, namely: providers, end-users, net visitors, and the nature of the traffic, whether institutional or individual-based. In all quarters, the global reception to this burgeoning technology has been remarkable. One illustrative example of this trend is OpenAI's ChatGPT. According to statistics, within five days of its launch in November 2022, the product had accumulated one million users [2]. By January 2023, this figure increased to an impressive 100 million users. By the mid-point of 2023, specifically June, the platform recorded an excess of 1.6 billion visits. Moreover, by January 2024, the visitor count to the ChatGPT website had surged to a staggering 2.3 billion, which included 2 million developers that were integrating its application programming interface (API). This exponential growth underscores the increasing incorporation and dependence on AI by both individuals and institutions, indicative of the transformative role of generative AI in the contemporary digital context [3].

Further testament to the shifting outlook towards generative AI is encapsulated in a report by McKinsey, dated March 2024 [4]. According to the report, the initial wave of enthusiasm and heightened activity in 2023 has evolved into a phase of readjustment, as companies grapple with the challenges of harnessing the potential value of generative AI. Business value derived from generative AI has been noticeably trending higher in 2024. A McKinsey Global Survey on AI conducted on May 30, 2024, indicates that about 65 percent of respondents confirmed their organizations' regular usage of generative AI [5]. This signifies a nearly twofold increase from the proportion recorded in the preceding survey conducted ten months prior.

Among the users of generative AI are thousands of companies of all size and industrial/service sectors, including those in healthcare or law, where a heightened privacy concern exists.[1] Initially showcasing as a translator or search engine on steroids, the latest advances, as advertised by AutoGPT, reveal a versatility that can create entire websites, conduct market research, and automate complex multistage physical or analytical operations without close human supervision. During the past year, open-sourced generative AI models, such as GPT4All, Dolly, Vicuna, llama.cpp, Ghostwriter, and PrivateGPT run as closed systems on desktops or laptops. This latter capability only serves to accelerate the diffusion of this innovation for the general user and companies motivated to remain competitive.

No society, profession, service or organization will escape the effects of generative AI technology, including those who by virtue of choice or circumstance remain bystanders to this technological shift.

II. Technology

Unlike machine learning AI, which typically deals with a closed set of subjects or topics, e.g., classifying objects, such as required by facial recognition systems, or controlling a self-driving automobile, a generative AI is capable of creating new content. LLMs began solely as language translators. The capability of generative AI to create new content was a consequence of an invention called the transformer (the "T" in GPT) to convert LLM's ability to translate any output content responsive to a set of input statements. In fact, the heart of the LLM is the transformer, which is a type of artificial neural network ("ANN"), developed, in 2017, by a team at Google Brain, to process sequential data inherent in any language or data time-series [6]. The following year, 2018, a team at OpenAI reportedly pre-trained a large-scale transformer using a text data corpus obtained from both public data and "data licensed from third-party providers"[2] [7]. The model was then fine-tuned through human reinforcement learning feedback [8]. These activities combined causing a quantum leap in natural language processing, which successfully demonstrated parallelization and training on extremely large datasets. Parallelization allows for efficient utilization of computational resources, such as computer processors that have multiple-cores with hyper-threading, or specialized graphical or tensor processors that operate at the rate of trillions of flops, reducing the time required for training.

In the context of neural networks, "parameters" are numerical values derived from data during the training process, which are referred to as weights and biases that a model applies to neurons or nodes. These parameters enable the model to generate content, such as language statements or images. The GPT-3 model currently consists of 175 billion parameters and utilizes 96 layers of neural networks used to compare the input to patterns learned from training data. OpenAI has not thus far revealed the number of parameters utilized in the GPT-4 model, but some would put the number in the realm of 1.76 trillion parameters, across 120 layers, which makes it 10 times larger than GPT-3.

A transformer in and of itself comprises a type of neural network designed for sequentially processing tokens, such as key parts of a words used in a sentence, or data points

[1] In the U.S. there no overarching federal privacy law governs data protection on the Internet, although several states have implemented their own privacy laws. Specific matters of privacy may be protected under the Privacy Act of 1974, which protects U.S. citizens from the misuse of their data by federal government agencies but is not applicable to private parties. Other federal laws, such as the Health Insurance Portability and Accountability Act (HIPAA), the Children's Online Privacy Protection Act (COPPA), and the Gramm-Leach-Bliley Act, specifically protect the rights and data of U.S. consumers, patients, minors, generally. A few state laws regulate the handling of Internet data, including data breach notifications, consent requirements, and consumer rights. Examples include the California Consumer Privacy Act (CCPA) and the Virginia Consumer Data Protection Act (CDPA).

[2] Microsoft holds a 49% stake in OpenAI, and collectively another 49% is held by a16z, Sequoia, Tiger Global, and Founders Fund.

used in a sequence or string. Each token is first embedded into a high-dimensional vector space, which allows the model to compare and mathematically manipulate the tokens in meaningful ways. Part of this process importantly includes a self-attention mechanism allowing the transformer to capture relationships between different tokens in the input sequence. For each token, the model calculates a weighted sum of the other tokens in the sequence where weights are determined by the similarity between the tokens. This allows the model to focus on the most relevant parts of an input sequence for each token, and enables it to capture complex dependencies between the tokens.

Transformer layers include feedforward networks, which process data and pass results to subsequent layers. Part of the computation process includes estimating the conditional probabilities of generating a word (based on its token) given the previous word or sequence of words to predict the next likely word or sequence of words based on context. For example, in the self-attention mechanism, the probabilities, i.e., the weights, control how much attention each word should pay to other words in the input sequence.

Although techniques like attention maps can provide some insights into which parts of the input text the model focused on, they do not provide a complete explanation of the decision-making process, because it utilizes statistical processes. And while it is possible to analyze the input-output relationship to gain some appreciation of an LLM's behavior, knowing precisely how its neural network determines a particular result is a practical impossibly, a point I shall elaborate on below as it factors into the potential for producing output having unintended consequences.

III. Performance

Developers of a GPT class product claim their product can take the written portion of the U.S. multistate bar exam (MBE) and the Graduate Record Exam (GRE) General Test, reaching performance levels upwards of 75% and 90% respectively, representative of successful test taker's scores [9]. When supplied a biochemical molecule, the generative AI product turns-on its biochemical expertise to produce variations of the molecule. Generative AI also threatens to swallow a chunk of what was once an exclusively human-inspired domain: composition, art, the production of media, and political persuasion. Concern abounds about the potential of these systems to replace skilled workers in the performance of certain tasks, which now require considerable training, e.g., of artists in the production of graphic arts, of copywriters engaged in the production of advertisements or entertainment. It is also capable of creating new tools for programmers and "do it yourself" (DIY) nonprogrammers, tasked with writing software.

For decades AI has been part of our lives operating below the surface, outside our spheres of attention. More recently our attention has been drawn to AI's power to drive social media content, determine access to credit, determine access to health care, to predict election results, stock performance, weather, and sporting outcomes, and to provide gambling strategies. It is becoming a central feature in hiring decisions, biomedical analysis, medical device operation, robotic surgery, driverless vehicles, and in piloting airplanes [10], [11], [12]. In all respects, modern life runs into AI at some level.

The myriad of suppliers of generative AI systems are of two general kinds. There are those that develop tech advancements to the systems that provide the engines for multi-modal synthesizing like transforming one creative domain into another, e.g., text into either text, images, videos, or functional computer code. Secondly, there are those that address specific topics, where outputs are analytical or diagnostic, and that provide insights or advice as might be helpful in diagnosing a medical disease from an abnormality appearing in an MRI, CT scan or X-ray.

Generative AI, such as ChatGPT, already has had an impact in medicine. A recent study assessed the performance of GPT-4 in diagnosing complex medical cases. It compared GPT-4's success rate to that of medical-journal readers. Results showed that GPT-4 correctly diagnosed 57% of cases, outperforming 99.98% of simulated human readers generated from online answers.

While these results highlight the potential of generative AI to be a powerful supportive tool for medical diagnosis, further improvements, validation, and addressing ethical considerations are needed before clinical implementation. More will be said about this in Section IV, below.

GPT4All allows integration into commercial products, while other models, such as those based on Meta's Llama, are limited to non-commercial research uses only. Incidentally, ample training information is available for those moderately fluent in programming to create a DIY product for the home or the office.

Lawyers develop legal practice skills and resources over time. As they become rooted in a practice area, they naturally resort to techniques and resources that work-well in carrying out an assignment. Contract negotiation tends to be in this category. Spellbook, a GPT-4 law-directed product, analyzes and suggests contract language using a Microsoft Word environment. It ensures privacy by encrypting data throughout the entire process. Because GPT has quick access to an enormous collection of relevant, and often new and effective information, it puts negotiations between parties on a plane different from that currently practiced. As the technology of Lexis/Nexus profoundly changed the legal profession's approach to research, these kinds of products will change law practice as it utilizes generative AI to prepare legal briefs, to better curate legal decisions, and to estimate with greater accuracy which party stands the greater likelihood of prevailing on an issue.

As the state-of-the-art currently exists, a GPT-4 output already strikingly matches human-level performance in limited domains, exhibiting signs of Artificial General Intelligence (AGI) [13], [14]. However, OpenAI, the company responsible for the development of ChatGPT-4, has opted to keep the exact number of parameters utilized in the model's training under wraps [15]. A calculation by AX Semantics, another leading entity in automated content production, suggests that the figure could be as staggering as 100 trillion [16]. This colossal number, as per AX Semantics' interpretation, brings

the language model's functionality significantly closer to the sophisticated language and logic processing abilities of the human brain [17].

In May 2024, OpenAI launched GPT-4o (omni) a multimodal AI that accepts, generates and outputs content across text, audio, image, and video mediums [18]. GPT-4o's response speed to audio inputs is at a near-human level at an average of 320 milliseconds, which serves to elevate conversational abilities. Unlike earlier versions, such as GPT-4, GPT-4o can sense and respond to multiple speakers, tone, and background noises, outputting human-like vocalizations such as laughter, singing and emotion. The AI's capacity to handle over 25,000 words supports long-form content and sustained conversations. GPT-4o claims enhanced reasoning compared to prior products enabling it to tackle an expanded array of complex problems as well as to engage in real-time collaboration with users involving editing and iteration for both creative and technical writing tasks.

Generative AI technology is said to have the potential to advance to an AGI level of performance in the next 5 to 20 years, depending on the prognosticator. AGI is a hypothetical technology that would perform cognitive-like tasks at or above the level of a human. Such cutting-edge technologies do not currently exist but according to thought leaders Jensen Huang (CEO of NVIDIA) and Altman (CEO of OpenAI), AGI might be realized within the next five years, while Dario Amodei (Co-founder and CEO of Anthropic) predicts "human-level" AI in two to three years [19]. In the largest survey of its kind, 2,778 researchers who had published in top-tier artificial intelligence venues were queried for their predictions on the pace of AI progress and impacts. Forecasters give "at least a 50% chance of AI systems achieving several milestones by 2028, including autonomously constructing a payment processing site from scratch, creating a song indistinguishable from a new song by a popular musician, and autonomously downloading and fine-tuning a large language model." The survey reports that, "If science continues undisrupted, the chance of unaided machines outperforming humans in every possible task was estimated at 10% by 2027, and 50% by 2047. The latter estimate is 13 years earlier than that reached in a similar survey we conducted only one year earlier$\cdots$ " [20].

Generative AI already has showcased qualities that suggest the elusive machine equivalent of human cognitive abilities. Thus, despite the uncertain forecasts of AGI's advent, its inevitability, by itself, should serve as a wake-up call to hasten the development of policy and projects to foster responsible innovation into the unforeseeable future [21], [22], [23]. Until now, commentators, educators, elected officials, and government bureaucracies have largely ignored AI and its implications to sway public opinion. But we ignore its power to persuade at our peril as this paradigm shift will precipitate a reorganization of society on multiple levels, not the least which involves employment, education, justice, medicine, invention, science and government. Every product has an inherent life cycle, reliability as to eventual failure, and intrinsic limitations, which are a consequence of the laws of physics, chemistry and computation. In this regard, any product may not be completely controllable by humans, institutions or technological systems when it comes to the limits of its performance or safety.

IV. Innate Limits Prevent Absolute Assurances

The governance of emerging science and innovation poses significant challenges for contemporary democracies, at all levels, both individual and institutional [24]. Generative AI is no exception in that it can propagate misinformation, spam, and phishing; abuse legal and governmental process; and abet such practices as fraudulent academic writing, legal arguments based on bogus authority, produce fake images that demean and discredit individuals for economic, social and political advantage and produce fake identification documents. One example occurred in May 2023, where a deceptive AI-generated dystopian political advertisement was released in U.S. by the Republican National Committee, offering a glimpse into how the latest AI tech could be used in this year's U.S. election cycle. The ad prompted Congressmen Yvette Clarke (D-NY) to introduce a bill to require disclosures of AI-generated content in political ads. The power of generative AI technology will undoubtedly advance in accuracy and usability, but also become less discoverable in its deceptive intents.[3] The current capability for detecting deepfakes is weak to non-existent. Security experts appear to offer little advice beyond recommending that greater investments be forthcoming in more sophisticated deepfake detection technologies [25]. Some of these would include improved identity verification systems, including the use of biometric and liveness verification to prevent the misuse of deepfakes in identity theft.

As an ethical and legal principle, purveyors of LLM products must ensure that products are helpful, honest, and harmless. As Yampolskiy reminds us, "The unprecedented progress in artificial intelligence (AI) over the last decade, came alongside multiple AI failures and cases of dual use causing a realization that it is not sufficient to create highly capable machines, but that it is even more important to make sure that intelligent machines are beneficial for humanity [26]."

For decades, ethicists have been warning that AI raises the capability claim and the value claim. The capability claim speaks to AI system performance, e.g., how fast it computes, its accuracy and the complexity of tasks it can achieve. It also considers whether AI is or can become sufficiently capable of inflicting major damage to the commercial, private and governmental sectors within its reach.

The assertion pertaining to alignment emphasizes the obligation of an AI system to operate as an advantageous and accountable entity in preserving the complex mosaic of human values and morality across a broad array of backgrounds [27].

[3]Recently there has been a steady stream of fakes as for instance, when the social media site X was bombarded with AI-generated porn of Taylor Swift, https://www.theverge.com/2024/1/25/24050334/x-twitter-taylor-swift-ai-fake-images-trending and in other venues when Tom Hanks was falsely shown promoting dental scams, https://news.bloomberglaw.com/litigation/mrbeast-tom-hanks-stung-by-ai-scams-as-law-rushes-to-keep-pace?utm_source=substack&utm_medium=email; or Emma Watson's voice was duplicated for reading Mein Kampf, https://www.vice.com/en/article/dy7mww/ai-voice-firm-4chan-celebrity-voices-emma-watson-joe-rogan-elevenlabs?utm_source=substack&utm_medium=email. (Accessed Mar. 23, 2024).

Conversely, the subject of alignment also deals with the risk that AI systems pose in exploiting loopholes or taking shortcuts in ways inconsistent with ethical precepts. These phenomena are commonly referred to as "reward hacking" or "specification gaming."

As AI systems advance and tend toward greater autonomy of operation the likelihood for instability or noncompliant functionality vis-à-vis alignment increases. Unlike closed systems such as ChatGPT or Gemini, which training data cut-off dates are 2021 and 2023 respectively, other products, such as Prompt-to-OS (P2OS), Grammarly, ProWritingAid, Descript and Lumen are not closed systems, and may employ content from interactions with professional services and users to improve their model's performance [28]. Any generative AI could theoretically be trained on any dataset including ongoing discussions with one's clients, patients, or customers [29]. A widely accepted premise in the development of any technological system holds that as "feedbacks [positive or negative become more complex, so does the achievement of stability become more difficult and the likelihood of instability greater" [30]. This may be especially true for self-generative technologies such as generative AI, which has the potential to train or innovate without human supervision, and thus potentially compromise a prescribed alignment.

Since the inception of the modern-day computer, it has been widely accepted that a software program can replicate itself [31]. If that program comprises a neural network, it has the prospect for evolving as it replicates inculcated traits (e.g., copy weights and parameters) gathered from a former generation [32], [33]. Recently researchers reportedly succeeded in developing a hypernetwork that predicts the parameters for the new network in fractions of a second, which in theory could make transformer training unnecessary [34]. AI developed in this way may modify its performance, or improve its efficiency or accuracy but unconstrained and without supervision may also produce outcomes that are increasingly unstable respecting the alignment of normative ethical and moral values.

Concerns specific to AI and AGI are well-reported, notably as applied to AI driven autonomous weaponry, the production of BOTs, or the deployment of malicious code [35]. In light of these kinds of applications, generative AI based applications should be limited in their ability to threaten value and capability claims, especially via the production of computer code.

Major AI developers are working to address present and potential safety issues, although clearly a plethora of complex problems remain unsolved, unsolvable and yet to emerge. To this end generative AI developers have ongoing programs to develop methods that encode desirable AI behavior in simple and transparent forms, as well as informing our understanding and evaluation of AI decision making [36], [37].

Google, typical of the larger companies creating generative AI technology, claims it is committed to collaboration and safeguards in carrying out responsible development and addressing risks as AI becomes more capable. In its December 2023 product release of Gemini, Google states, "Gemini has the most comprehensive safety evaluations of any Google AI model to date, including for bias and toxicity. We've conducted novel research into potential risk areas like cyber-offense, persuasion and autonomy, and have applied Google Research's best-in-class adversarial testing techniques to help identify critical safety issues in advance of Gemini's deployment [38]."

In December 2023, OpenAI announced its Preparedness Framework, to tackle risks posed by cyberattacks or autonomous weaponry, via consistent risk evaluations, predefined safety measures, and continual capability assessments of performance of products, such as ChatGPT [39]. The framework also strives to spot and handle emerging risks through data-based predictions. Although these efforts appear to indicate a concerted and conscientious approach to AI deployment and advancement, in May 2024, OpenAI dismissed a team of researchers specifically organized to work on mitigating AI misuse, economic disruption, disinformation, bias and discrimination, addiction and overreliance. While it is essential to recognize that organizational decisions can be complex and multifaceted, this development may signal a period of transition or reevaluation, and thus that OpenAI is in state of flux in attending to these important issues [40].

A product developer's commitment to a technology's underlying innerworkings or transparency is a hallmark of responsible innovation. However, as relates to generative AI an inescapable reality must be confronted. Generative AI technologies (such as GPT-3 and GPT-4) employ neural networks that make decisions stochastically, i.e., based on parameters and user inputs, and thus by design, do not permit a complete understanding about how a decision is made. An assessment of its unintended ramifications to health, safety and welfare cannot be analytically determined.

In the past, many technologies have been suspected of being harmful to humans or the environment. In some cases, it took decades to understand how a technology adversely affected health, such as regarding cigarette smoking or pesticide exposure, but over time science was able to establish the causal connections between a population-wide application of what were physical products and their effect on health. In respect to non-physical, that is, intangible products, such as generative AI technology, a calculation produces an expression that humans interpret as information. This type of computation/interpretive cause and effect has an ontologically different characterization compared to the cause-and-effect phenomena that manifest between physical objects. The effect of a chemical on one's physical or psychological condition cannot be framed in the same way as the effect that information has on one's physical or psychological condition. In short, they are different things and the consequences are different, the first being instantiated in physics and chemistry and the other instantiated in a social construction, which is replete with cultural, political and economic implications.

While LLMs can produce impressive results, they lack explicit understanding or reasoning about the content they generate. Generative AI does not explicitly store information or provide step-by-step reasoning for its outputs. As stated, a model's calculation is distributed across a complex neural network, making it a practical impossibility to predict any specific output. Although researchers may eventually understand

the process by which an output chooses a sequence of words, there will remain an uncertainty as to the precise words chosen [41]. By analogy we may understand how a pair of dice functions to generate a numbered pair, but we cannot predict in advance what pair of numbers will be revealed in the next toss. As such, neural networks generally pose special problems in risk assessment as the path through which data propagates, as weighted by the parameters established through training, cannot be determined. Thus, the risk appertaining to the unknowable issue results from the statistical nature of a generative AI model, limiting our understanding of how it arrives at specific output.

There are risks associated with particular kinds of inventions when their inner workings cannot be fully understood. This characteristic as applied to AI driven products like generative AI has become a field unto itself under the rubric "safety and security" [42]. The innate lack of transparency, in the model's decision-making process, necessitates that users exercise caution and critical evaluation when using generative AI outputs in sensitive or high-stakes contexts. Companies such as Anthropic and OpenAI as well as other researchers are actively working on developing methods to improve interpretability and explainability of AI models, but it remains an ongoing challenge [43].

A lack of understanding and imagination on the part of legislators as to how the technology works and its potential to do harm in the general sense, will limit the effectiveness of any proposed regulation. Technologies such as nuclear power, were understood to have obvious devastating consequences as was demonstrated in 1945, and thus it required relatively little incentive to subject their use to strict regulation. But countless technologies exist that do not immediately manifest their potential for planet altering effects. Fossil fuels, cancer causing chemicals, and social media serve as examples that initially were obscured by a lack of knowledge about their potential to do harm. And often when the potential harms a technology is capable of visiting upon a population become apparent, policymakers tend to ignore the problem because of self-interest, political or otherwise, or due to large-scale skepticism about whether actions, or even warnings are necessary. Examples are: cigarette smoking, which was found to cause cancer; excessive fossil fuel use, which contributes to climate-change; and in the U.S., the reluctance of many to use masks during the recent COVID-19 pandemic; and the use of assault rifles in the commission of senseless mass murders.

Generative AI will alter how we apply technological power at all levels of society, especially those that depend on high-tech for their lifeblood. Presuming regulators begin to investigate this technology, they need advice from experts in various fields, such as computer science, technology, medicine, social science, economics, intellectual property, and ethics. In the first instance experts will define and specify the various failure modes, e.g., what might harm the public in a particular instance. It is important to not overly constrain the development of the technology, or limit or burden its distribution or application. Nevertheless, the public is entitled to a thorough understanding of the ways in which the technology could harm vital interests in health, safety, welfare, self-determination, and creativity.

The design stage of product development presents the most opportune time to mitigate the likelihood of a future errant operation due to a design flaw or potential ethical mishap. Included in this Special Issue is a paper entitled: "How to Regulate Large Language Models for Responsible AI," by Jose Berengueres. To achieve responsible AI in the context of LLMs, the author identifies touchpoints such as: (1) conducting a review of codes of ethics, (2) making an assessment of ethics awareness, and (3) identifying safeguard application points. Each of these touchpoints is then evaluated on the basis of cost and effectiveness. The key finding of the paper is that applying safeguards upstream aligns with established engineering practices in addressing issues at the source.

In the initial exploration stage of developing a generative AI model, a critical component that should be included would be a thorough analysis of potential failure modes within the particular product. This inquiry would endeavor to unveil possible malfunctioning channels, their origins, and their potential consequences. It is important to recognize that the use of the phrase "failure mode" in the case of generative AI technology means the applications to which the product conforms, and as consequence has the potential to proximately cause harm as a result of, e.g., inaccuracy, infringement of intellectual property, its vulnerability to cybercrime, intrusions to personal/individual privacy, its inability to explain its operation, abetting violations of equity and fairness, defamation, physical or psychological harm.

In specific cases such as a medical image analysis, the AI may fail to perform its intended diagnostic function. In a copyright infringement case, the failure mode may be its inability to detect plagiarism in the output. In a case where racial bias is alleged, the failure mode manifests in the inability to detect a bias present in the calculations and decision-making logic that leads to unequal treatment and outright discrimination. Here is partial list of questions, most of which are familiar to the engineering community, but may help guide a policy analysis of product pitfalls. Answering the following questions may help gain a deeper understanding of failure modes, thus allowing for the development of policies and strategies to prevent, detect, and mitigate harm in the development of products that utilize as its underlying system generative AI technology.

1) What are the intended functions and performance requirements for the generative AI product under design?
2) What are the possible failure modes that could occur during the generative AI product's lifecycle?
3) What are the anticipated causes or factors that could lead to each failure mode?
4) What are the impacts of each failure mode for the generative AI product, on users, or a particular application?
5) What is the probability of occurrence of each failure mode, and what is the severity of the consequences that flow?

6) Are there any safeguards or preventive measures in place to mitigate or prevent an identified failure mode?
7) Can a failure mode be anticipated and then detected or monitored through any means (system detection sensors, expert evaluations, tester or user panels, checklists, inspections, etc.)?
8) What are the potential warning signs or indicators that a failure mode is imminent?
9) What are the possible actions or countermeasures that can be taken to prevent, detect, or mitigate each failure mode?
10) How can the generative AI product design or pre-production testing, design review or quality control processes be improved to eliminate or minimize failure modes?

V. Responsible Innovation

Responsible innovation aims to maximize positive social and economic benefits while minimizing any unforeseen negative outcomes. The process includes an active effort to remove obstacles to its adoption and diffusion. The crucial questions for which we seek answers include: Are the benefits of a science or technology distributed evenly? How can we align technology and innovation with societal needs? And in any given case, what are an innovator's ethical responsibilities [44]?

In "Developing a Framework for Responsible Innovation," Stilgoe et al., address four dimensions of responsible innovation:

1. Anticipation: Foresight and proactive consideration of potential impacts and risks;

2. Reflexivity: Ongoing reflection and critical assessment of the innovation process;

3. Inclusion: Engaging diverse stakeholders and considering their perspectives;

4. Responsiveness: Adaptability and willingness to adjust based on feedback and changing circumstances [45].

These principles are designed to ensure that advancements are in harmony with ethical standards, societal requirements, and environmental sustainability. They prompt innovators to be reflective about their work, to engage with relevant parties, and to take the lead in fostering fair, inclusive, and socially sensitive institutional environments. Yet, no guarantee exists that these ideals will be forthcoming without formalizing processes that ensure they become part of the fabric of institutions where the work is carried on. An example of efforts to address this point, Erik Fisher a researcher for Socio-Technical Integration Research (STIR) heads a project at Arizona State University, which focuses on how science and engineering work in labs affects society. It recognizes that science and technology policies worldwide place new pressures on laboratories to consider broader societal implications. The project investigates how laboratories can respond to these pressures and the role that interdisciplinary collaborations play in responding to socio-technological integration, by providing an experimental platform for scientists and engineers to incorporate social science and humanities perspectives in the course of conducting their normal work. It is vital that socio-technological integration perspectives filter into the development of generative AI platforms in ways that serve to align AI outcomes to values of importance to humanity, values often found wanting in the narrow technological specifications of our projects.

The social, political and economic power of innovation is not just in its creation, but in its diffusion. Technology diffusion is "... the process by which innovations are adopted by a population. Whether diffusion occurs and the rate at which it occurs is dependent on several factors including the nature and quality of the innovation, how information about the innovation is communicated, and the characteristics of the population into which it is introduced. ... [46]" In the context of the generative AI products, adoption itself acts as a positive feedback, further conditioning and accelerating greater levels of adoption within specific social demographics, scenarios and frameworks. The rate of diffusion or integration of a new technology begins with individual understanding of the innovation, the strength of the argument about its claimed benefits, and the decision to embrace or reject the innovation [47]. This process also includes the practical use of the innovation, and its eventual solidification into a stable part of one's life or occupational practice. In many cases, individuals employ an innovation, because it is advantageous to a competitive position, or fulfills a desire to create utilitarian and nonutilitarian products or expressions, or it improves performance in the exercise of one's profession, such as academic, law, medicine, engineering, business, politics or the arts. The way we handle and adapt to these changes often defines success in personal affairs and in our occupations.

Generative AI is anticipated to expand its functional capabilities, but today, as a free-standing product its technological form appears to reliably perform as intended by its designers. Because of its revolutionary capabilities it has garnered an historic world-wide rate of diffusion, as mentioned earlier in Section I. It is unlikely that as currently configured there would be little that engineers and programmers might do to dramatically throttle its performance in any particular direction. Before us is a future, and we can only speculate as to the trajectory of this technology. Thus, what particular invention will spawn from the present AI technology is impossible to foresee with any degree of clarity.

Importantly, what needs further investigation and analysis is to what ends do we apply this technology, i.e., for what purpose. Generative AI as a breakthrough invention represents a component, analogous to a gear, or an instrument, like a typewriter [48]. Aside from the potential for innovation as to the component or instrument, our attention must be drawn to how we purpose or employ generative AI. In other words, what kind of applications should be regulated by government. This is a question that will need careful study as any course of action will have to surmount various cultural, political, economic, commercial and ethical points of view. Each point of view will have constituencies from every walk of intellectual, creative and social life.

VI. Legal Issues Anticipated to Surface in a World of Generative AI

As emphasized, without understanding how a technology works, it is impossible to identify its universe of harmful or socially objectionable performance permutations. For instance, although we may reduce the occurrence of threats caused by the production of provably untrue information, we cannot entirely eliminate it. This applies to the impossibility of reliably determining in advance if a particular application will produce an output that conforms to a society's normative capability and value claim. As mentioned, generative AI's stochastic underpinnings prevent a complete understanding of causation, which obfuscates and complicates the assignment of responsibility in cases of negative outcomes, creating legal, ethical, and regulatory challenges regarding accountability of liability and remediation [49].

A complete analysis of the range of legal causes of action and their likely outcomes is beyond the scope of this paper, as such would require considering actual cases. However, the following five areas address the typical but most salient kinds harm that might reasonably flow from various uses one might envision in generative AI applications based on widely litigated cases over the last several decades.

A. Privacy and Data Protection

Generative AI technology relies on vast amounts of data, which it acquires from various databases. It has been established that training data often includes text from sources, which may contain sensitive, private or proprietary information. If not properly anonymized or stripped of personally identifiable information, the use of this data in training generative AI models can result in unintentional exposure of private information.

Related to the privacy and data protection category are such matters as "posing and exposing." Because generative AI models use and create text resembling human language, the output can potentially lead to adverse inferences about matters private or confidential. For example, if a user interacts with a chatbot and provides personal details or discusses sensitive topics, a risk exists that the model might generate responses that at a future point in time indirectly reveal or expose that information. In a parallel context, it is plausible that generative AI may produce, or inadvertently amplify and disseminate, erroneous or deceptive information concerning an individual or an organization. This could be potentially harmful, defamatory, or intrusive to privacy. This may also encompass the manifestation of deepfake content within deceptive emails or in public domains such as social media, print circulation or even within advertisements.

A plaintiff in the U.S., who claims to have been harmed in connection with a data processing system breach, a generative AI product in this case, must typically resort to state common law causes of action and remedies. A few exceptions exist in cases where a state statute may cover computer crimes or torts where a computer is implicated in the harm. Another exception, for which state statutory causes of action exist relates to unfair and deceptive practices in a commercial context.

Federal agencies are liable for privacy breaches under The Privacy Act of 1974, where it grants a civil cause of action against the government related to privacy and data protection.[4] Other non-statutory right to privacy causes of action are recognized in 30 states and the District of Columbia, including: Trespass, which protects against physical intrusions or surveillance; Breach of Confidence, which addresses unauthorized disclosure of private information; Defamation, libel, or false light, which pertain to false statements or images that harm reputation; Appropriation of name or likeness; Intrusion upon Seclusion, which focuses on intentional invasions of privacy; and Misuse of Private Information or proprietary information. Related to proprietary and confidential information are violations of various trade secret statutes, which can be actionable under state or federal statutes and common law.

B. Copyright Infringement

Technology, creative content, and law join during the training of the system and then subsequently in the generated output. The following examples claiming copyright protection illustrate this point.

In Silverman v. OpenAI Inc., case number 3:23-cv-03416, a U.S. district judge dismissed most of the infringement claims against OpenAI and Meta. The court found that the authors, including Sarah Silverman, Michael Chabon, and Paul Tremblay, failed to demonstrate that the outputs of ChatGPT were "substantially similar" to the copyrighted books in question. This suggests a high bar that plaintiffs will need to overcome in proving an alleged infringement of a generative AI's output. The use of training data in the first instance is an entirely different matter.

Training data for generative AI products often comes from extensive datasets, which, despite including publicly available text, inevitably consist of copyrighted material. For instance, these generative AI datasets also use data scraped from online content to create "Scraping and Shadow Libraries," some of which are not segregated from copyrighted materials and/or escape filtering and copyright detection mechanisms. The unauthorized storage of such copyrighted material on any computer system is typically considered an act of infringement.

Defenses against infringement, such as Fair Use and Transformative Works, have also been lodged. These defenses, derived from statutory and case law over the past century, may or may not hold up due to the intricacies of an ultimate evidence-based analysis, which requires individual case evaluation.[5]

[4]The Privacy Act of 1974, known as Public Law No. 93-579, was enacted on December 31, 1974. It is codified at 5 U.S.C. § 552a and became effective on September 27, 1975. This law serves as the principal framework governing the handling of personal information within the federal government, emphasizing fair information practices and safeguarding individuals' privacy rights.

[5]Section 107 of the Copyright Act provides the statutory framework for determining whether something is a fair use. It identifies certain types of uses—such as criticism, comment, news reporting, teaching, scholarship, and research—as examples of activities that may qualify as fair use.

Concerning text generation, copyright infringement could potentially occur when AI-generated text unintentionally mirrors existing works. As a result, generative AI products must be carefully designed to prevent accidental plagiarism. Moreover, the fleeting nature of the content used for training could also serve as a defense. Because of the textual construction of a generative AI output, it might be impossible to discover evidence and thus ascertain that the original information used in an output is "substantially similar" to the original.

However, the absence of filtering and copyright detection mechanisms could potentially indicate that an infringement is deemed reckless or intentional. While copyright infringement is based on a strict liability standard, damages are often increased if the defendants were found to be willful in their actions.

Sora, the developing generative artificial intelligence model from OpenAI, excels in creating short video clips using text, and breathes new life into the debate over generative AI's creative ownership. Traditionally, copyright has implicitly belonged to human authors. However, the rise of Sora presents a perversion to this precedent. When Stephen Thaler attempted to register a copyright, the work was rejected on grounds that artificial intelligence, as the creator of content, fell short of the requirement for human created enlightenment, personal authorship, unique creativity, and discretionary choice, all considered necessary for copyright protection. In response, Thaler sued the Copyright Office, which led to a court ruling in August 2023 [50]. The court's decision sided with the defendant, asserting that U.S. copyright law only recognizes humans as eligible authors [51]. Consequently, the AI-produced work did not qualify for copyright protection. The decision was supported by a policy statement published by the Copyright Office in March 2023, in which AI-composed content was distinctly ruled out for copyright protection [52]. Sora could be viewed as a remarkable leap forward in the world of AI content creation. Nevertheless, it will continue to provoke legal disputes over the definition of authorship, and how an initiator of a creative endeavor might protect the intellectual property primarily and autonomously produced by software.

C. Bias, Fairness

Trainers of generative AI are in a position to decide which pieces of information to incorporate into training, as for instance, by using pre-training feedback that helps frame outputs in consideration of societal norms. Their choices, like those of any human curator, influence the depth, excellence, and impartiality of the information landscape. Whether consciously or subconsciously, designers bring their own viewpoints, ethics, and standards about the world to the table.

Nevertheless, generative AI systems can and do inherit biases from the data they are trained on, leading to unfair outcomes or discrimination [53]. Generative AI is typically trained using databases that originate in what might be referred to as the Western world. As such the databases comprise information having an inherent cultural bias and outputs will reflect that fact. One would anticipate that if the training used databases from Asia, Africa or the Indian Subcontinent, output would reflect biases pertaining to those cultures. How these problems may manifest are impossible to predict, as pointed out by Theodore McCullough, who contributes to this Special Issue, in his paper "Explaining and Exploring Ethical and Trustworthy AI in the Context of Reinforcement Learning." The author cites the case of a Kenyan Sama employee who asked OpenAI an earnest question as a labeler as how to tag a piece of sexual content, distasteful from a Kenyan point of view, but perhaps condoned otherwise by a Westerner [54]. This open-ended example illustrates the challenge and complexity of societal norms that differ across the globe. As to the present state of pre-trained generative AI, the matter of cultural diversity appears intractably burdened by biases, which have the potential to engender consequences without a remedy.

A possibility for ensuring that diverse mores, traditions and ways of life are incorporated into generative AI products may be found in a requirement that training incorporates information originating by and through different cultures. The scarcity of well-accumulated and fine-tuned multilingual text databases until now has posed challenges in this regard. Progress to rectify this deficiency have been reported with the release of the CulturaX dataset—a compilation of 6.3 trillion tokens across 167 languages. This kind of addition to the family of LLM systems may serve to improve the multilingual capabilities of extensive language models, expanding their linguistic and cultural range [55].

But other more common types of bias can and do exist in AI used in hiring decisions, health care access and many other activities encountered in life. This can be avoided and remediated technologically and enforced through laws dealing with the harm that computer AI systems often perpetrate and perpetuate.

There are many examples to draw upon, but a 2019 report serves a type of algorithmic bias that has life and death implications [56], [57]. The algorithm, sold by leading health services company, Optum, had been shown to dramatically underestimate the health needs of mostly unwell African American patients, thereby reinforcing the long-standing racial disparities within the medical field. The algorithm utilized by Optum influenced the decision-making process for the healthcare of millions. However, researchers suggested that this problem was likely prevalent in other tools employed by different private companies, nonprofit health systems, and government agencies. Rectifying this bias would more than double the number of African American patients identified as being at risk of complex medical needs within the health system studied. Researchers have collaborated with Optum to develop a solution. Upon replicating the analysis on a national dataset of 3.7 million patients, it was found that African American patients, who were ranked by the algorithm as being in equal need of additional care as white patients, were actually significantly more ill. Unfortunately, the law as developed does not satisfactorily remedy what an individual may have suffered as a consequence of this set of unfortunate circumstances.

D. Safety and Reliability

Certain generative AI applications may find their way into applications, such as counseling or medical diagnosis and will have direct impact on human lives, which will require safety standards, testing requirements, and liability frameworks to ensure that these systems are reliable, trustworthy, and do not pose unnecessary risks. In part this was covered generally in Section IV above, and will be covered further in Part E, below, as to how product liability law may prove a cause of action in matters related to malpractice claims. Medical devices, which now may include generative AI that will aid physicians in treatment and diagnosis, should and may well be required to follow Federal Drug Administration (FDA) rules and regulations[6] [58]. However, algorithmic decision-making tools used in clinical, administrative, and public health settings — such as those that predict risk of mortality, likelihood of readmission, and in-home care needs risk falling outside current regulations.

E. Product Liability

Multimodal AI is an advanced form of artificial intelligence that can understand and create content across various modalities such as text, images, and audio simultaneously. In the future it will undoubtedly be employed in innovations that exploit its power to control mechanical and electrical apparatuses and processes. The technology may act both as a control mechanism and a programmed source of advice, thereby positioning it under the category of products as per the law. In the realm of product liability law, a product is deemed dangerous if it presents a risk of harm to consumers, either due to inherent flaws or defects. This area of law is often connected with the concept of strict liability. This concept implies that the manufacturers or sellers can be held responsible for any injuries caused by their products, irrespective of their intent or knowledge about the defect. This is distinct from negligence, which demands proof of fault. Strict liability, in contrast, is primarily concerned with the safety and condition of the product.

Under product liability law, defects can be classified into different categories. In the case of generative AI, the defect is most likely termed a design flaw, making it inherently risky, even when manufactured correctly. Product liability litigation necessitates that a plaintiff demonstrate that the product was defective when it left the defendant's possession and that it was the cause of the plaintiff's injury. One of the well-established judicial standards, known as the "consumer expectation standard," posits that a product is flawed if its potential danger is both unknowable and unacceptable to an ordinary consumer.

The onus of analytical transparency must not be solely on the developers. The users, mainly professionals who employ generative AI devices in their practice, also have an ethical and legal duty to scrutinize the information derived from generative AI through systematic vigilance of its integrity. Consider the medical imaging community, for instance. They significantly benefit from generative AI's advanced algorithms, which operate within the AI system's framework, and are now experiencing a marked enhancement in the precision of the analysis of medical images. This is resulting in earlier disease detection, reduced diagnostic errors, and better patient outcomes. However, in situations where due diligence is required, physicians cannot absolve themselves of blame or liability because a machine caused the errant result [59].

In the legal arena, lawyers bear a similar duty of care to ensure that their advice, legal documents, and representation are accurate. While the professionals themselves carry the ethical and legal responsibility to their clients and the court, the organizations that oversee these professionals have an obligation to ensure that educational and compliance measures align with the principle of "do no harm." They must also exercise increased vigilance when leveraging resources utilizing generative AI.

Overall, regulations on the use of generative AI through the administrative and judicial branches of government, should strive for a balance between fostering innovation and protecting societal interests, ensuring that the related technologies are developed and deployed to benefit humanity. Unfortunately, evidence to date suggests that we cannot depend on the U.S. government to alone counter or mollify the adverse consequences of generative AI technology. Chief concerns, about a timely government response, stem from: (1) the speed with which the generative AI technology is diffusing throughout society, and (2) its power to generate content, both in the non-utilitarian space, such as art, prose, or poetry, and in the utilitarian space of computer code, or invention in the traditional sense of machines and processes, and (3) the inability to comprehend the computational parameters and pathways the technology utilizes in achieving an output.

VII. Governmental Regulation

The developed world is rife with unbridled commercialization, fierce competition, and political instability, each country through its high-tech establishment pushing the boundaries of technological conquest. However, with advances in know-how comes responsibility. Powerful tools in the hands of irresponsible agents always threaten the fabric of civilization. Query: will governments heed the warnings and work alongside developers in an effort to advance humane goals, or simply allow AI technology to propagate unconstrained by the value, to "do no harm?"

There have been recent efforts in both Europe and the U.S. to reign in AI initiatives. In October 2022, the White House Office of Science and Technology Policy published The Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. This is an example of a generalized omnibus type governance applicable to all AI designs and deployment. It does not mandate U.S. policy but states principles by which the government and industry might find footing. It is discretionary as to its adoption. Other

[6]"The Food and Drug Administration (FDA) has long regulated software that meets the definition of a device in section 201(h) of the Federal Food, Drug, and Cosmetic Act (FD&C Act), including software that is intended to provide decision support for the diagnosis, treatment, prevention, cure, or mitigation of diseases or other conditions (often referred to as clinical decision support software)."

examples of preliminary action are the Department of Defense AI Ethical Principles and Responsible AI Implementation Pathway and the Intelligence Community AI Ethics Principles and Framework and The National AI Initiative Act of 2020, which became law on January 1, 2021. These kinds of overarching principles often gain traction in private industry when it is required that they are adopted as a condition of being awarded government contracts.

In September 2022, the FDA passed regulations dealing with the use of AI in medical devices. Its guidance states that some AI tools should be regulated as medical devices as part of the agency's oversight of clinical decision support software. The guidance includes a list of AI tools that should be regulated as medical devices, including devices to predict sepsis, identify patient deterioration, forecast heart failure hospitalizations, and flag patients who may be addicted to opioids. The FDA recognizes that AI and machine learning particularly have been increasingly incorporated into medical devices because these algorithms are capable of "learning" from experience and improving performance over time.

Over the course of history, the U.S. has established numerous regulatory agencies to deal with technology. For example, the Federal Communications Commission (FCC) and the FDA regulate communications, drugs, and medical devices. But when it comes to the more amorphous forms of digital technology, such as data gathering, data security or the reach of the Internet, regulation has remained lethargic. The government has yet to regulate any concrete aspect of social media.

To successfully regulate any technology requires experts in the technology and its application. This has been true for communications as well as medical technology. For example, as to drugs the FDA enlists chemists, physicians, statisticians, patients, and policy experts to effectively regulate. The Select Committee on AI, created in June 2018, advises the White House on interagency AI R&D priorities. Within the last six months, the Executive Branch appears to be in the early stages of assembling meaningful AI oversight commissions.

In October 2023, President Biden issued Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [60]. Broad policy and principles were promulgated for AI development and use. Agencies were ordered to adhere to these principles albeit considering input from various stakeholders. According to the order, future initiatives must address safety and security, requiring robust evaluations, standardized testing, and risk mitigation, as related to biotechnology, cybersecurity, and critical infrastructure.

The administration followed up on the earlier Executive Order in April 2024, reporting that during the intervening 180 days federal agencies had addressed a broad range of AI's safety and security risks, including those related to dangerous biological materials, critical infrastructure, and software vulnerabilities [61]. The actions included establishing a framework for nucleic acid synthesis screening to prevent the misuse of AI for engineering dangerous biological materials, the release of draft documents for public comment on managing generative AI risks, securely developing generative AI systems, expanding international standards development in AI, and reducing risks posed by AI-generated content. Importantly, an AI Safety and Security Board has been formed, with 22 members to advise the Secretary of Homeland Security, the critical infrastructure community, other private sector stakeholders, and the broader public on the safe and secure development and deployment of AI technology in critical infrastructure.

In spring 2023, the U.S. Senate Committee on the Judiciary, Subcommittee, Artificial Intelligence and Human Rights convened hearings to investigate the general concerns AI poses [62]. Since its last meeting in June 2023, there has been no reported progress from the U.S. Congress, and given the current dysfunction in the legislative branch, and the breadth of diverse commercial interests at stake, it may be that attempts to regulate generative AI technology via legislation will not succeed in having any measurable impact for the foreseeable future.

In considering the broad questions regarding responsible AI technology generally, the U.S. National Institute of Standards and Technology (NIST) inaugurated the U.S. Artificial Intelligence Safety Institute (USAISI) on November 1, 2023. Operating under its auspices, a collective of over 200 organizations has been assembled to devise guidelines and standards for AI measurement and policy. These are rooted in scientific methodologies and empirical evidence, serving as the bedrock for global AI safety. This policy endeavors to assist the U.S. in managing the potential risks associated with future AI models and systems. Ranging from cutting-edge models to novel applications and strategies, the Institute is prepared to address and navigate the evolving landscape of artificial intelligence [63]. Perhaps most promising is that the consortium of institutions forming the core of the USAISI are themselves research and development actors in AI technology. We might anticipate that members will import their participation in the consortium into their roles as corporate managers, and advocate within their respective industries for policies and practices consistent with responsible AI innovation. Notwithstanding the USAISI's aims, there remains the necessity to regulate the bounds of safe and effective AI through legislation and action by agencies charged with ensuring that their respective codes of federal regulation are enforced.

Specific to AI, the European Commission has identified applications of AI based on their potential for widespread harm and has moved to install the European AI Act (AI Act), which addresses risks of specific uses of AI. The AI Act applies to AI machine learning, expert and logic systems, and Bayesian or statistical approaches whose outputs "influence the environments they interact with," which includes generative AI products such as ChatGPT. The legislation distinguishes four categories of AI use: unacceptable AI risk, high-risk, limited risk, and minimal or no risk.

On March 13, 2024, a pivotal shift in the AI landscape, undoubtedly felt internationally, was marked by the European Parliament's adoption of the AI Act [64]. The AI Act, designed to safeguard fundamental rights and democratic principles from high-risk AI applications, also aims to stimulate innovation in the AI sector. The European Council unanimously approved the EU AI Act on May 21, 2024, which is anticipated to be officially published, mid-year 2024. After a

span of 24 months following publication, the AI Act will be fully enforceable, with specific clauses activated on different timelines.

The AI Act squarely outlines obligations for systems based on their potential risk and impact levels, banning AI applications that could infringe on individual rights. This includes the use of biometric classification, indiscriminate facial recognition, emotion recognition in workplaces and educational institutions, social scoring, predictive policing, and AI that manipulates human behavior or capitalizes on vulnerabilities. Law enforcement's usage of biometric identification systems has been constrained, except in strictly defined situations. Real-time deployment is subject to specific authorization, such as in targeted searches for missing individuals or in preventing terrorist activities. Post-incident use, referred to as "post-remote biometric identification," requires judicial authorization related to criminal offenses [65].

High-risk AI systems, specifically those with significant potential to harm health, safety, fundamental rights, environment, and democracy, are subjected to rigorous regulation. A key requirement referred to as the Fundamental Rights Impact Assessment (FRIA) requires deployers of high-risk AI systems, including public bodies and certain private operators, to assess potential impacts on fundamental rights, such as privacy and non-discrimination, before deploying these systems [66]. This includes the assessment of risk and transparency, maintaining usage logs, guaranteeing accuracy, and ensuring human oversight. The AI Act provides the right to lodge complaints against entities that exploit AI systems in a manner that violates rights under the act. Furthermore, AI systems must adhere to EU copyright law and disclose summaries of training data. More stringent measures are stated for AI models that could pose systemic risks, necessitating model evaluations, risk assessments, incident reporting, and labeling of artificial or manipulated content.

It appears that generative AI has the potential to sense and express abstractions and human motives, as well as create code, self-learn, and therefore lead to the instantiation of the technology into robots, broadly speaking, which may exhibit autonomous behavior. As mentioned, these applications likely face regulation under the AI Act, which prohibits employing untargeted facial images for facial recognition as well as banning emotion recognition in workplaces and schools.

Earlier, the European Parliament resolution of 16 February 2017 offered recommendations to the Commission on Civil Law Rules on Robotics. This was particularly aimed at the creators, manufacturers, and users of robots that were equipped with self-learning capabilities and inherent autonomy. According to this directive, they were expected to adhere to Asimov's Laws to ensure a robot did not act in a manner detrimental to human interests. This necessitates that any generative AI system would need to implement safeguarding protocols to prevent the creation of code that would violate this fundamental principle. Currently generative AI technology does not explicitly prevent such code being incorporated into an output.

The ratification of the AI Act is a crucial milestone in Europe's strategy to regulate AI, striking a delicate balance between fostering innovation and protecting fundamental rights in the digital age. The AI Act may reduce the potential for the deleterious impacts of generative AI on entire populations by ensuring that in extreme cases, e.g., autonomous weapons or medical devices, developers of the technology will be subjected to a measure of jurisprudential scrutiny and control. The AI Act will also raise compliance issues with U.S. companies, when their AI products inevitably cross a European border. As such, U.S. companies that develop, deploy manufacturer, export, or distribute AI systems for use in the European Union, will be required to adapt or harmonize the operation of their generative AI products so as to conform to the requirements under the Act. The Act's procedures aside, the technological underpinnings and safeguards of all products, regardless of where they are utilized, will therefore likely conform to the European standard [67].

VIII. Conclusion

Generative AI technology amplifies human ingenuity. The invention will prove to benefit science and the humanities into the far reaches of time. But inventions of this magnitude do not come free from misuse, harm and obligation. It is anticipated that the technology will substantively infiltrate all sectors of the society, affecting the economy, as well as the health, safety and welfare of its citizens, and potentially the institution of democracy itself. Needless to say, the ease of use and the seemingly unbounded breadth of generative AI applications will spawn a range of important innovations having economic, political, ethical, and legal ramifications.

We should take comfort in the fact that AI's power and success will not stop humans from composing, authoring, or inventing, as we are wired to express ourselves in ways that ensure our survival, both materially and aesthetically. Yet, generative AI will foster new inventions. Some of these will take form in utilitarian products and processes, such as new article of manufacture, apparatuses, compositions of matter and others will manifest in music, art and authorship. Still others will take form in non-utilitarian objects, such as AI generated human-like avatars, posing as actors, hucksters, and politicians, or as humanoid robots for companionship and commercial utility [68].

We cannot ignore that coupled with the human contribution to generative AI products and processes, societal change will dwarf the kind of transitions the world experienced going from horse-driven carts to high-speed autos, bull horns to television, or snail mail to email.

The potential for generative AI technology to change the social, commercial and creative landscape calls for an initiative to carefully consider the value claim and the capability claim as to whether these products will always act according to human values, such as "do no harm," which are aligned with those of humanity, and if not whether its actions could cause significant harm. Efforts in both the U.S. and Europe to regulate AI applications have been observed, such as the U.S. FDA's regulation of AI in medical devices and the European AI Act categorizing AI uses and requirements for oversight based on potential harm. In the U.S., over time, courts will adjudicate and thus clarify rights, duties and obligations that developers and users owe to their communities

and constituents. But the power of generative AI technology coupled with its rapid diffusion into society-at-large, requires a comprehensive plan of oversight and collaboration between government, AI developers and those who will commercialize LLM related products and services. In the U.S., it is incumbent upon Congress and the Executive Branch to heed the warnings and proactively initiate regulation through a new commission to work alongside developers and companies that intend to market LLM applications, to ensure that AI is advanced and controlled in a manner that aligns with humane goals and avoids potential harm. The assertion that humans are extrinsic to, or merely peripheral to the operation of AI neglects the indisputable reality that humans themselves moved to commence this inventive construct. And thus, only humans can possess the will and the ability to comprehend the significance of the events that transpire within such a construct [69].

REFERENCES

[1] (Ethics Centre, Sydney, NSW, Australia). *Ethics Explainer: Double-Effect Theory*. Mar. 29, 2016. [Online]. Available: https://ethics.org.au/ethics-explainer-double-effect-theory/

[2] "Number of ChatGPT users." May 2024. Accessed: Jun. 3, 2024. [Online]. Available: https://explodingtopics.com/blog/chatgpt-users

[3] "107 up-to-date ChatGPT statistics & user numbers." Mar. 2024. Accessed: Jun. 6, 2024. [Online]. Available: https://nerdynav.com/chatgpt-statistics/

[4] "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value." May 30, 2024. Accessed: Jun. 3, 2024. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[5] "A generative AI reset: Rewiring to turn potential into value in 2024." Mar. 4, 2024. Accessed Jun. 3, 2024. [Online]. Available: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/a-generative-ai-reset-rewiring-to-turn-potential-into-value-in-2024#/

[6] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010. [Online]. Available: https://doi.org/10.48550/arXiv.1706.03762

[7] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, Jun. 11, 2018, "Improving language understanding by generative pre-training," Dataset. Accessed: Mar. 18, 2023. [Online]. Available: https://paperswithcode.com/paper/improving-language-understanding-by

[8] J. Achiam, "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[9] "Model card and evaluations for claude models." 2023. Accessed: May 10, 2024. [Online]. Available: https://www-cdn.anthropic.com/files/4zrzovbb/website/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226.pdf

[10] J. R. Carvalko Jr., "Future of pharmaco-electronic medicine," *Quinnipiac Health Law J.*, vol. 25, no. 3, pp. 355–490, 2022.

[11] D. Camacho, M. V. Luzón, and E. Cambria, "New research methods & algorithms in social network analysis," *Future Gener. Comput. Syst.*, vol. 114, pp. 290–293, Jan. 2021. [Online]. Available: https://doi.org/10.1016/j.future.2020.08.006

[12] J. Park, M. Cheon, S. Hou, and O. Lee, "Forecasting election result via artificial intelligence approach: nlp and machine learning," in *Proc. Int. Conf. Commun. Comput. Technol.*, 2023, pp. 761–768. [Online]. Available: https://doi.org/10.1007/978-981-19-3951-8_57

[13] S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," 2023, *arXiv:2303.12712*.

[14] B. Edwards. "OpenAI's GPT-4 exhibits 'human-level performance' on professional benchmarks: For another accounting of the GPT-4 model performance." Ars Technica. Mar. 14, 2023. Accessed: May 9, 2024. [Online]. Available: https://arstechnica.com/information-technology/2023/03/openai-announces-gpt-4-its-next-generation-ai-language-model/

[15] E. Griffith. "GPT-4 vs. ChatGPT-3.5: What's the difference?" PCmag. Mar. 16, 2023. Accessed: Mar. 23, 2024. [Online]. Available: https://www.pcmag.com/news/the-new-chatgpt-what-you-get-with-gpt-4-vs-gpt-35 A. Geyer, "What you need to know and what's different from GPT-3 and ChatGPT." Semantics. Mar. 13, 2023. Accessed: May 9, 2024. [Online]. Available: https://www.ax-semantics.com/en/blog/gpt-4-and-whats-different-from-gpt-3

[16] Ibid.

[17] Ibid.

[18] "Hello GPT-4o, We're announcing GPT-4o, our new flagship model that can reason across audio, vision, and text in real time." May 13, 2024. Accessed: May 15, 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/

[19] A. McFarland. "Could we achieve AGI within 5 years? NVIDIA's CEO Jensen Huang believes it's possible." Unite AI. Mar. 18, 2024. [Online]. Available: https://www.unite.ai/could-we-achieve-agi-within-5-years-nvidias-ceo-jensen-huang-believes-its-possible/ W. Henshall. "When might AI outsmart us? It depends who you ask." Time. Jan. 19, 2024. Accessed: May 15, 2024. [Online]. Available: https://time.com/6556168/when-ai-outsmart-humans/

[20] K. Grace, H. Stewart, J. F. Sandkühler, S. Thomas, B. Weinstein-Raun, and J. Brauner, "Thousands of AI authors on the future of AI," 2024, *arXiv:2401.02843*.

[21] B. Perrigo. "U.S. must move 'decisively' to avert 'extinction-level' threat from AI, government-commissioned report says." Time. Mar. 11, 2024. Accessed: Mar. 15, 2024. [Online]. Available: https://time.com/6898967/ai-extinction-national-security-risks-report/

[22] S. Altman. "Planning for AGI and beyond." OpenAI. Feb. 24, 2023. Accessed: Mar. 23, 2024. [Online]. Available: https://openai.com/blog/planning-for-agi-and-beyond

[23] J. Leike and I. Sutskever. "Introducing superalignment." OpenAI. Jul. 5, 2023. Accessed: Mar. 23, 2024. [Online]. Available: https://openai.com/blog/introducing-superalignment

[24] J. Stilgoe, R. Owen, and P. Macnaghten, "Developing a framework for responsible innovation," *Res. Policy*, vol. 42, no. 9, pp. 1568–1580, 2013.

[25] G. Bueermann and N. Perucica (World Econ. Forum, Cologny, Switzerland). *How Can We Combat the Worrying Rise in the Use of Deepfakes in Cybercrime?*. May 19, 2023. Accessed: Mar. 23, 2024. [Online]. Available: https://www.weforum.org/agenda/2023/05/how-can-we-combat-the-worrying-rise-in-deepfake-content

[26] R. V. Yampolskiy, "On the controllability of artificial intelligence: An analysis of limitations," *J. Cyber Secur. Mob.*, vol. 11, no. 3, pp. 321–404, 2022, doi: 10.13052/jcsm2245-1439.1132.

[27] E. Nabavi, R. Nicholls, and G. Roussos, "Locating responsibility in the future of human–AI interactions," *IEEE Trans. Technol. Soc.*, vol. 5, no. 1, pp. 58–60, Mar. 2024, doi: 10.1109/TTS.2024.3386247.

[28] G. Tolomei, C. Campagnano, F. Silvestri, and G. Trappolini, "Prompt-to-OS (P2OS): Revolutionizing operating systems and human-computer interaction with integrated AI generative models," 2023, *arXiv:2310.04875*.

[29] P. Li, "Announcing new capabilities for azure OpenAI on your data." May 21, 2024. Accessed: Jun. 5, 2024. [Online]. Available: https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/announcing-new-capabilities-for-azure-openai-on-your-data/ba-p/4144636

[30] W. Ross Ashby, "Feedback, adaptation and stability, selected passages from design for a brain (the origin of adaptive behaviour)." 1960. [Online]. Available: https://www.panarchy.org/ashby/adaptation.1960.html

[31] J. von Neumann, "The general and logical theory of automata," *Collected Works*, vol. 5, A. H. Taub Ed. New York, NY, USA: Pergamo, 1956, pp. 288–328. M. J. E. Golay, "Reflections of a communications engineer," *Anal. Chem.*, vol. 33, no. 7, Jun. 1961. p. Bratley and J. Millo, "Computer recreations: Self-reproducing automata," *Softw., Pract. Exp.*, vol. 2, no. 4, pp. 397–400, 1972.

[32] K. Mok, "AI researchers create self-replicating neural network." The New Stack. 2018. [Online]. Available: https://thenewstack.io/ai-researchers-create-self-replicating-neural-network/

[33] J. Carvalko, *The Techno-Human Shell: A Jump in the Evolutionary Gap*. Mechanicsburg, PA, USA: Sunbury, 2013, pp. 95–97.

[34] A. Ananthaswamy (Quanta Mag., New York, NY, USA). *Researchers Build AI That Builds AI*. Jan. 25, 2022. Accessed: May 10, 2024. [Online]. Available: https://www.quantamagazine.org/researchers-build-ai-that-builds-ai-20220125/

[35] W. Wallach, *A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control*. New York, NY, USA: Basic Books, 2015.

[36] E. Perez et al., "Discovering language model behaviors with model-written evaluations," 2022. *arXiv:2212.09251*.

[37] "Claude's constitution." Anthropic. May 9, 2023. Accessed: Aug. 7, 2023. [Online]. Available: https://www.anthropic.com/index/claudes-constitution

[38] S. Pichai and D. Hassabis, "Introducing Gemini: our largest and most capable AI model." Dec. 6, 2023. Accessed: May 15, 2024. [Online]. Available: https://blog.google/technology/ai/google-gemini-ai/

[39] R. Banfield. "OpenAI outlines 'preparedness framework' to systematically track and mitigate AI safety risks." Maginative. Dec. 18, 2023. Accessed: May 15, 2024. [Online]. Available: https://www.maginative.com/article/openai-outlines-preparedness-framework-to-systematically-track-and-mitigate-ai-safety-risks/

[40] W. Knight. "OpenAI's long-term AI risk team has disbanded," Wired. May 17, 2024. Accessed: Jun. 2024. [Online]. Available: https://www.wired.com/story/openai-superalignment-team-disbanded/

[41] K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt, "Interpretability in the wild: A circuit for indirect object identification in GPT-2 small," 2022, *arXiv:2211.00593*.

[42] R. Yampolskiy, *Artificial Intelligence Safety and Security*, New York, NY, USA: Chapman Hall, 2018.

[43] J. R. Schoenherr, R. Abbas, K. Michael, P. Rivas, and T. D. Anderson, "Designing AI using a human-centered approach: Explainability and accuracy toward trustworthiness," *IEEE Trans. Technol. Soc.* vol. 4, no. 1, pp. 9–23, Mar. 2023, doi: 10.1109/TTS.2023.3257627.

[44] "What is responsible innovation?" UCL. Accessed: Mar. 23, 2024. [Online]. Available: https://www.ucl.ac.uk/responsible-innovation/what-responsible-innovation

[45] Ibid.

[46] (Boston Univ. School Public Health, Boston, MA, USA). *Technology Diffusion*. Accessed: Mar. 23, 2024. [Online]. Available: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/HPM/AmericanHealthCare_Technology-Drugs/AmericanHealthCare_Technology-Drugs2.html

[47] E. M. Rogers, *Diffusion of Innovations*, 5th Ed. New York, NY, USA: Simon Schuster, 2003.

[48] E. Braun, *Wayward Technology*, London, U.K.: Frances Pinter, 1984, p. 42.

[49] J. R. Schoenherr and R. Thomson, "When AI fails, who do we blame? Attributing responsibility in human–AI interactions," *IEEE Trans. Technol. Soc.*, vol. 5, no. 1, pp. 61–70, Mar. 2024, doi: 10.1109/TTS.2024.3370095.

[50] T. v. Perlmutter, "Memorandum opinion 2023 WL 5333236 (D.D.C.)," U.S. Copyright Off., Washington, DC, USA, document No. CV 22-1564 (BAH), Aug. 2023.

[51] Ibid.

[52] "Copyright office 37 CFR part 202 copyright registration guidance: Works containing material generated by artificial intelligence," U.S. Copyright Off., Washington, DC, USA, document 2023-05321, Mar. 10, 2023. [Online]. Available: https://public-inspection.federalregister.gov/2023-05321.pdf

[53] S. Akter et al., "Algorithmic bias in data-driven innovation in the age of AI," *Int. J. Inf. Manag.*, vol. 60, Oct. 2021, Art. no. 102387.

[54] B. Perrigo. "Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic." Accessed: Apr. 13, 2024. [Online]. Available: https://time.com/6247678/openai-chatgpt-kenyaworkers/ H. Guinness. "How does ChatGPT work?" Accessed: Apr. 13, 2024. [Online]. Available: https://zapier.com/blog/how-does-chatgpt-work/

[55] T. Nguyen et al., "CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages," 2023, *arXiv:2309.09400*.

[56] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31649194/

[57] C. Y. Johnson. "Racial bias in a medical algorithm favors white patients over black patients." Washington Post. Oct. 24, 2019. Accessed: Mar. 23, 2024. [Online]. Available: https://www.washingtonpost.com/health/2019/10/24/racial-bias-medical-algorithm-favors-white-patients-over-sicker-black-patients/

[58] (Food Drug Admin., Silver Spring, MD, USA). *Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff*. Sep. 28, 2022. [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software

[59] Ibid, 27,49,69.

[60] "Executive order on the safe, secure, and trustworthy development and use of artificial intelligence." Oct. 30, 2023. [Online]. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[61] "Biden-Harris administration announces key AI actions 180 days following president Biden's landmark executive order." Apr. 29, 2024. [Online]. Available: https://www.whitehouse.gov/briefing-room/statements-releases/2024/04/29/biden-harris-administration-announces-key-ai-actions-180-days-following-president-bidens-landmark-executive-order/

[62] (U.S. Senate Comm. Judic., Washington, DC, USA). *Artificial Intelligence and Human Rights*. Jun. 13, 2023. Accessed: Mar. 23, 2024. [Online]. Available: https://www.judiciary.senate.gov/committee-activity/hearings/artificial-intelligence-and-human-rights

[63] (U.S. Dept. Comm., Washington, DC, USA). *At the Direction of President Biden, Department of Commerce to Establish U.S. Artificial Intelligence Safety Institute to Lead Efforts on AI Safety*. Accessed: Mar. 23, 2024. [Online]. Available: https://www.commerce.gov/news/press-releases/2023/11/direction-president-biden-department-commerce-establish-us-artificial

[64] "Artificial intelligence act (COM(2021)0206—C9-0146/2021—2021/0106(COD))," Eur. Parliam., Strasbourg, France, Rep. A9-0188/2023, Accessed: Mar. 23, 2024. [Online]. Available: https://www.europarl.europa.eu/doceo/document/A-9-2023-0188-AM-808-808_EN.pdf

[65] Ibid.

[66] H. Waem, J. Dauzier, and M. Demirican. "Fundamental rights impact assessments under the EU AI Act: Who, what and how?" Technology's Legal Edge. Mar. 7, 2024. Accessed: May 23, 2024. [Online]. Available: https://www.technologyslegaledge.com/2024/03/fundamental-rights-impact-assessments-under-the-eu-ai-act-who-what-and-how/

[67] V. Peaden, "EU AI regulation ripples through tech value chains, U.S. business," Bloomberg Law. Mar. 21, 2024. Accessed: May 13, 2024. [Online]. Available: https://news.bloomberglaw.com/us-law-week/eu-ai-regulation-ripples-through-tech-value-chains-us-business

[68] Á. F. Gambín, A. Yazidi, A. Vasilakos, H. Haugerud, and Y. Djenouri, "Deepfakes: Current and future trends," *Artif. Intell. Rev.*, vol. 57, p. 64, Feb. 2024. [Online]. Available: https://doi.org/10.1007/s10462-023-10679-x

[69] K. Michael, J. R. Schoenherr, and K. M. Vogel, "Failures in the loop: Human leadership in AI-based decision-making," *IEEE Trans. Technol. Soc.*, vol. 5, no. 1, pp. 2–13, Mar. 2024, doi: 10.1109/TTS.2024.3378587.

Joseph R. Carvalko Jr. (Member, IEEE) is a technologist, an academic, and a patent lawyer. In the course of his career, he had worked for over two decades in research and development as an engineer, a researcher, and a technician. He is a named inventor in 18 U.S. patents in various fields, including artificial intelligence. His law practice includes counseling, patent prosecution, and litigation mainly dealing with technology and intellectual property. He has authored academic books, articles, and fiction throughout his career. He is currently the Chairman of the Technology and Ethics Working Research Group, Interdisciplinary Center for Bioethics, Institution for Social and Policy Studies, Yale University; and is an Adjunct Professor of Law with Quinnipiac University, School of Law, teaching Law, Science, and Technology; a member of IEEE Society on Social Implications of Technology; and a member of the Publications Board of IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY. His latest book *Conserving Humanity at the Dawn of Posthuman Technology* provides an account of AI and genetics from a technical, historical, and ethical perspective as well as expectations for its future development. More about him can be found at https://carvalko.com/bio/.