ETHICAL AI
GOVERNANCE
GROUP

BEYOND THE BLACK BOX: Shaping a Responsible AI Landscape

2023

Dear Readers,

As we reflect on the transformative journey of artificial intelligence in our society, we are reminded of the profound moments that have shaped our digital age. The year 2023 marks what many are calling the "Netscape moment" for AI. Thanks to innovations like ChatGPT and its widespread adoption in B2C use cases, we have transitioned from viewing AI as a mysterious "Black Box" to recognizing its direct relevance to the everyday user. This evolution has ignited public interest and sparked discussions on a scale we've never seen before. AI is no longer confined to the silos of tech enthusiasts; it has become a topic of household conversation, for better or worse.

The winds of change have not only been felt in the realm of technology but have also resonated in the corridors of power. Europe, in its characteristic foresight, has taken the lead in AI regulation. This move has not only intensified the ongoing discourse but has also amplified the zeitgeist that invariably accompanies every major technological disruption. At its core, this regulatory shift is a testament to the undeniable fact that the very ground beneath our feet has transformed. We are not in the same world we once knew.

In this ever-evolving landscape, opinions on the opportunities and risks presented by AI are as diverse as they are passionate. It's easy to get swayed by the cacophony of voices, each asserting its version of the truth. However, in this annual report, we at EAIGG have made a conscious effort not to lean too heavily in one direction. Instead, we present to you a collection of thought-provoking articles from our esteemed members. Our aim is to elevate the discourse, offering insights that can illuminate the path forward and foster a platform for meaningful dialogue.

By incorporating perspectives from investors, policymakers, and enterprises, we believe this report offers a unique balance. More crucially, these viewpoints provide a practitioner's lens, shedding light on the tangible ways AI professionals are developing and deploying these technologies. Such grounded insights anchor our discussions where they truly belong – on solid, pragmatic ground.

In closing, I invite you to delve into these pages with an open mind. Let's embrace the complexities, celebrate the advancements, and together, chart a course for an AI-driven future that is ethical, inclusive, and truly transformative.

Warm regards,

Anik Bose

Anik Bose
Executive Director
EAIGG

A Special Thanks to KPMG for their support of the 2023 Annual Report

KPMG

TABLE OF CONTENTS

THE ‘NETSCAPE MOMENT:’ UNDERSTANDING THE IMPACT OF GENERATIVE AI

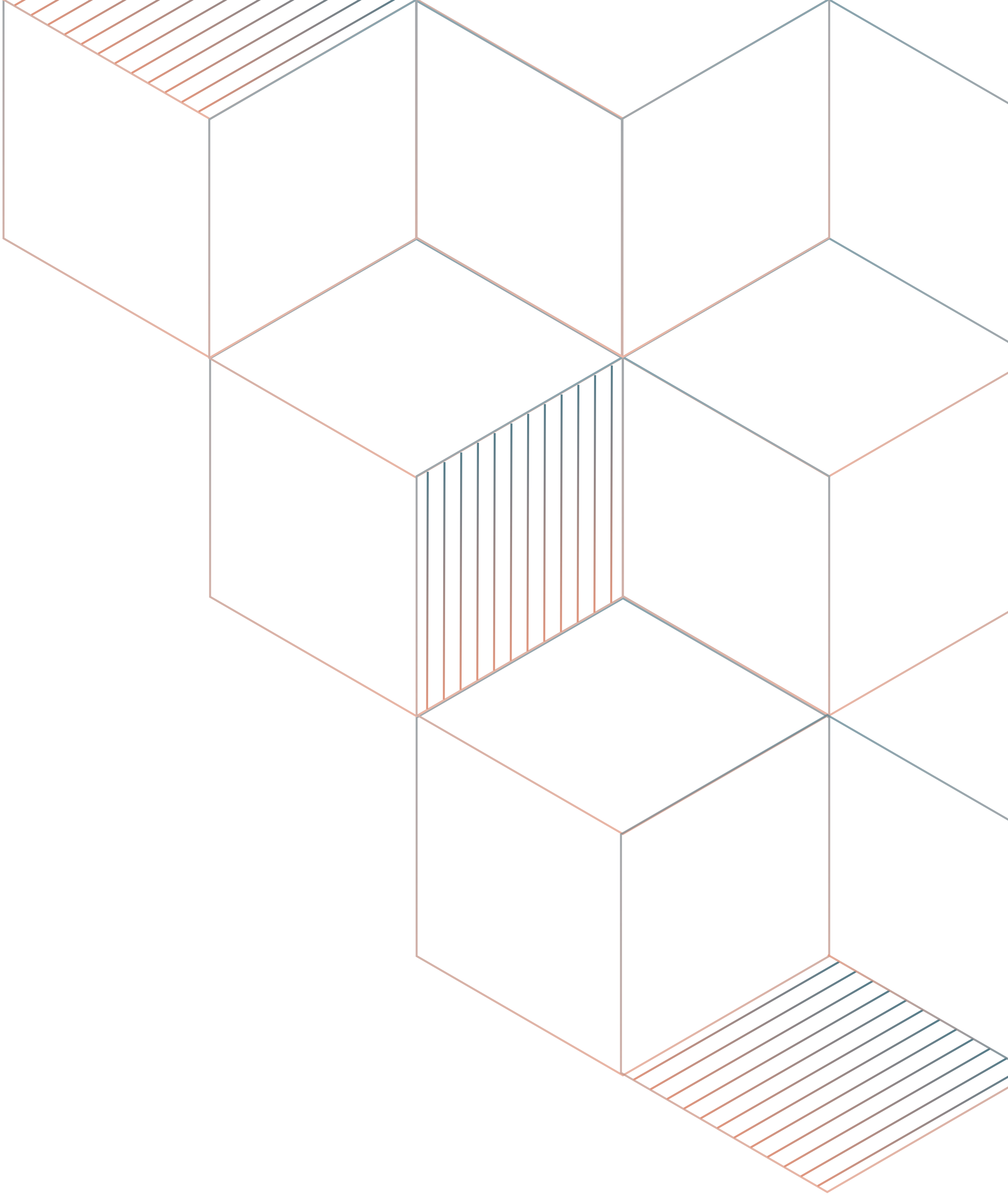
REGULATION IS COMING: ETHICAL POLICYMAKING IN PRACTICE

TABLE OF CONTENTS

INNOVATION AT THE FOREFRONT: UNLEASHING AI'S FULL POTENTIAL

INVESTING IN AN ETHICAL FUTURE: BUILDING AI RESPONSIBLY

CONCLUSION



THE ‘NETSCAPE MOMENT:’ UNDERSTANDING THE IMPACT OF GENERATIVE AI

Breaking Down the Basics: Understanding What Is and Is Not Generative AI

Jelmer Van Der Linde

Biography

Jelmer van der Linde is currently a Research Software Engineer at The University of Edinburgh, where he's a part of the machine translation group, focusing on projects like ParaCrawl. He earned his Bachelor of Science in Artificial Intelligence and his Master of Science in Artificial Intelligence from the University of Groningen, with an interest in subjects like natural language processing, computer vision, and autonomous systems. Before joining The University of Edinburgh, Jelmer was an API/Web Developer at Global Surface Intelligence for over 5 years, where he played a pivotal role in developing platforms for geospatial data. He also spent over 8 years at the University of Groningen as an Application Developer and Research & Teaching Assistant, guiding lab sessions across a variety of AI-related subjects. Additionally, Jelmer has a rich freelance history, having worked as a Web Developer for 15 years, collaborating on small-scale web projects. His commitment to technology was further showcased during his tenure at Concept7, where he built an inventory management system and migrated a blog to WordPress.

Breaking Down the Basics: Understanding What Is and Is Not Generative AI

Divyansh Agarwal

Biography

Divyansh is a Senior Research Engineer/Applied Scientist with the conversational AI team at Salesforce AI Research. His work involves developing and studying large-scale AI applications that allow real-world user interaction with AI models. He is active in Natural Language Processing (NLP) research on topics like text summarization, AI for search and robustness of large language models.

Divyansh graduated from the Language Technologies Institute (LTI) at Carnegie Mellon University (CMU) in Pittsburgh. He has been a fellow with Internet Society, actively advocating on topics like policy, ethics and multi-stakeholderism on the internet.

Divyansh is passionate about studying the interaction of humans with AI technologies, the interplay of ethics and privacy in that mix, and how that can inform the future generation of AI applications.

Breaking Down the Basics: Understanding What Is and Is Not Generative AI

By Jelmer van der Linde and Divyansh Agarwal

Generative Al is already the buzzword of 2023 in tech. The release of chatGPT by OpenAl has pushed Al out from the bubble of academics and a handful of industry practitioners, into the spotlight of mainstream media and the general populace. Tech giants have realized the game changing applications of this technology, and have been quick to respond with a barrage of new systems powered by Generative Al models in their bid to stay competitive. With new developments in this space making headlines almost every single day, it's important to break down and understand this concept, which will likely inform the future generation of digital systems people use.

To understand how these systems function at their core, it's imperative to understand how these large models were developed by Al scientists in the first place. At the same time, to understand how their performance is evaluated in the right direction, we need to be able to distinguish between the systems that are built on top of Generative Al models, and the models themselves. This in turn brings the process of developing these systems and the importance of human oversight to the fore.

What is Generative AI?

Generative Al is essentially a field of Al that deals with building (and studying) algorithms/ models trained to generate creative digital artifacts like text, images, videos etc. This generative ability is a result of repeatedly making the Al model learn to understand simple natural language prompts and the expected response it should produce. Generative Al did not have a Eureka moment out of the blue as it may seem but has been evolving for some time in the research community and the industry. It was years of incremental research that led to powerful image generation models like DALL-E and Stable Diffusion, along with the Al models that could generate realistic videos given a text prompt, that started coming out in 2021-2022. Similarly, it was several iterations of developing text generation AI models that ultimately led to the inception of ChatGPT, a technology which seems to be the defining moment for Al in general. Its astonishing performance was the tipping point for the tech industry to find this confidence (and billions of dollars in funding) in recognizing Generative Al as a game changer. This in turn has led to a new text generation model (like Claude, Bard, Vicuna etc.) being announced almost every single week.

The different flavors of Al models that generate text, images, videos and multimodal content, are equally impactful in their own respect. However, in order to get a better understanding of how these models are developed, let's dive deeper into text generation models like ChatGPT, which are more generally known as large language models (LLMs).

Deep Dive

Language Models (LMs) are AI models/algorithms that are trained to understand (and generate) natural language text. They have been studied for the past decade (and more) in the field of Al and Machine Learning. Training an LM boils down to teaching a model to implicitly understand word associations in natural language. For several iterations during its training, an LM is given some input text, and it gradually learns to generate the expected output word for word. Everyday applications like the auto complete which help write your emails, Google Translate or the chatbot that answers your questions, all use LMs under the hood. Once an LM is trained on a specific dataset, it can perform that task really well, like predicting, classifying, summarizing, translating text etc., in a specific language (but of course, multilingual models exist too!). When these LMs scale in terms of size, a.k.a Large Language Models, or LLMs, it essentially increases their capacity to learn multiple tasks simultaneously. What the LLM learns in one task benefits its performance on the other (related) tasks as well. One single LLM that has enough capacity, when trained using the right techniques and ample data, can then perform various generation tasks (like ChatGPT does).

But what did ChatGPT do differently to have such a good performance as an LLM? Is it the training technique, or the data? Well, both the modeling technique employed and the natural language text data contribute to the performance of an LLM. When it comes to the data, the trend follows the principle that the volume of the data itself is more important than the specific nature of it. (And ChatGPT indeed was trained on an unprecedented volume of data, spanning multiple tasks like question answering, text summarization etc.) However, the learning mechanisms for training these LLMs, is something that the AI research community has iteratively (but significantly) improved in the last few years. The focal point of this revolution in LLMs arguably came about in 2017, with the development of the transformer AI model by researchers at Google. In the next few years, a flurry of LLMs inspired by this modeling technique were developed by AI researchers, maturing progressively in size, function and outperforming existing benchmarks at breakneck speed. Iterative research led to the development of the techniques in AI that would involve humans in the loop and prompt based learning, while training these LLMs on massive datasets for several iterations. 'Reinforcement Learning from Human Feedback' (RLHF), the technique that powered previous models by Open AI, used a novel method to incorporate human feedback during the LLM training. Building on previous research in RLHF, and some careful changes employed in the training, was the magic sauce that led to the creation of ChatGPT, an LLM that outperformed previous models with astonishing accuracy. But does ChatGPT work all that well in all respects?

Image generated using DALL-E 2. Prompt: Programmer making generative Ai in a computer lab, digital cyber punk

The Unpredictability and Unreliability of LLMs

It is indisputable at this point that large language models are great at generating fluent and naturally sounding text, and can adapt to many different domains. Writing a professional sounding resignation letter for a fictional role, or a computing science tutorial as given by a pirate, are not a challenge. This is what it is trained to do: produce fluent output, and later with RLHF, produce believable and dare I say pleasing output. This doesn't even seem to be a complex task: a competitive English German machine translation system can store all the knowledge it needs, including grammar rules for both languages, in just 17 million parameters (measures of complexity of AI systems).

But if the number of parameters of the model is increased into the ranges of large language models (GPT-3 consists of 175 billion), and give it the training data to fit all those parameters, you end up with a model that can recall facts such as who is the ruling monarch in England, and seemingly even learn the rules to complex tasks such as arithmetic, or rhyme in poetry.

It is difficult to understand what rules the model learned exactly. For the model, these rules are merely patterns observed in examples, not the result of experimentation of rules it was told about. Its training objective is to predict the next word in the answer, and it learns that following these patterns is an effective way of doing that. Do not be fooled, this simple method is really powerful! ChatGPT for example, when prompted to do chain-of-thought reasoning, in which it writes out each of the intermediate reasoning steps, can use its own intermediate output to power-pattern-match its way to correct answers. For example, when asked to solve a math question using arithmetic, this method is quite effective (albeit horribly inefficient when compared to how a classic computer program would solve this. And when this question is altered in a way that it needs to be solved analytically, ChatGPT will output relevant analytical observations, and then appear to reason towards a conclusion. But as expected, when any of these observations are irrelevant or wrong, the conclusion is likely to be too. This is fuzzy pattern matching, not the infallible arithmetic we're used to from computers and calculators.

The same holds true for its knowledge about facts. We have to remember that all these stored tidbits are effectively a side effect of how the model is trained, where it learns to predict the correct answer word for word. All knowledge, whether it is grammar, semantics, or rules for reasoning, is learned in the same way. And we cannot attempt to alter one without possibly affecting the others. This is problematic if your facts change.

Even in these massive models, the knowledge is not exact. The model is a derived artifact from the data it was trained on. It is lossy compression, where knowledge that occurs more often in the training data is more likely to be preserved in great detail in the model. In a way, LLMs are like blurry jpegs, and the training data is not stored verbatim. As a result, the model cannot guarantee to be able to reproduce the exact source of a fact it mentioned: the data might not be there. Worse, it cannot tell whether it generated a fact as seen in its training data, or produced an amalgamation of different facts into a fictitious one. When you play with the newest version of ChatGPT, it will often produce a correct quote, name, title or url because that exact sequence has been prevalent enough in the training data that the model has learned that these are in themselves likely sequences. But, unlike a search index as used in a search engine, there is no guarantee: ChatGPT will produce fake headlines without any indication that they do not exist. And since the training data is also often not or only partially published, it can be tricky to verify whether ChatGPT answered true to its training data, or made something up.

These models are being used to perform multiple tasks, where the description of the task is given to the model as part of the prompt. Previously these instructions would have been expressed in code, which we know how to debug, and execute in a predictable manner. But with these models we rely on it following instructions. The big win here is that it is no longer needed to expertly design and implement complicated algorithms, performance improves by just training bigger models with more data. And when the model is based on a pre-trained LLM, like a GPT or LLAMA (an LLM released by Meta), the knowledge embedded in these is an amazing starting point. Without any specific training, these pretrained LLMs will likely be able to for example perform ROT13, a simple substitution cipher, without any specific training on how that cipher works. Just from the knowledge that was in the massive amounts of data the model was pre-trained on. But unlike executing code, language prompt answering is not exact. Slight variations in the input can produce radically different outputs. Out of domain input can yield completely unpredictable output. And even when a model is provided with instructions that describe an exact algorithm; the execution will be a (wordly) approximation that may get the details wrong. For example, when asking ChatGPT to perform ROT13, it will come close, but fumble some words. Even when the Wikipedia explanation of ROT13 is added to the prompt. This is interesting because ROT13 is not about words, it is about replacing each letter with another. Yet ChatGPT substitutes one word for a shorter or longer similar word. This highlights that a language model is not a calculator: you can give it instructions, but it is still trained to predict text. It might predict an execution of those instructions, but there is no guarantee. This also introduces an interesting new security risk when making a language model a part of a system: the instructions the model gets, and any input from the user, often come through the same channel, and it is possible for the model to be confused or be abused.

In short, LLMs can be unpredictable and unreliable. Slight variations of input can result in completely different outputs. Instructions can be ignored. And there is no distinction between fact and fiction.

Putting Generative AI systems into perspective

It's important to realize that generative AI models like ChatGPT are just a small component of the whole 'Generative AI tech stack'. As we collectively realize the endless applications of Generative AI models, we are inevitably moving towards a future where humans interact with AI systems more and more. Building end-to-end applications would require thinking beyond the Generative AI model to the systems themselves, that involve managing user data and interactions, along with the infrastructure, model lifecycle management etc. We are only just beginning to realize how these interactions are different from users communicating with traditional non-AI based systems.

When designing Generative AI systems, it's imperative to allow the user to have a sense of reliability, such that they feel that their interactions are dependable, secure and factually correct. Given the associated skepticism emanating from the budding nature of the field itself, sometimes one shot is all an AI system gets at building this trust with a user. As humans, we also find the need for Generative AI systems to clarify the source(s) of the information presented to us, in order to reliably put it to good effect. As Generative AI applications evolve, perhaps their greatest value is in personalizing content for a particular user, and catering to our diverse set of criterias and preferences as to what information is relevant. Not only are some of these factors vital design principles for Generative AI Systems, but they raise new and important questions for all of us to find answers to.

A case of responsible development of Generative AI

Generative AI is not new, but its general availability and sudden increase in capabilities is. There is a wide world of possible applications for these models, and an intense investment frenzy to get us there. Generative AI opens up new capabilities to humans, such as making professional looking art without needing years of experience holding a digital paintbrush, and envisions new possibilities in simplifying and innovating our technical systems. This in turn raises many ethical questions surrounding data, legality, the authenticity of derivative works, or who is responsible for a machine's output and its consequences. Grounding the development of Generative AI systems as a whole in ethics has never been more important. As the applications of this technology multiply, it is vital to have an ethics board as an important part of the Generative AI tech stack. In order to sustain the growth of Generative AI, and guide its impact in the right direction, responsible AI development practices should be employed by the industry and academia alike.

KEY TAKEAWAYS

- Generative Al is a field of Al that deals with building algorithms and models trained to generate digital artifacts like text, images and videos. This generative ability is reached by training an Al model to understand simple natural language prompts and produce the most likely response.

- Generative Al models and LLMs can be unpredictable and unreliable. They may give us a correct and relevant response, but they can also give us answers which are incorrect and irrelevant. Given that we're unable to identify the sources used by an LLM, it can be difficult to identify whether or not the output is fact or fiction. We will need to bear these shortfalls in mind when using these technologies.

- Generative Al is not new, but its general availability and sudden increase in capabilities is. As the applications and uses of generative Al multiply, so will the ethical questions surrounding data, legality, accountability, and authenticity. To guide the impact of generative Al in the right direction, we will need to employ responsible Al development practices and ground the development of generative Al as a whole in ethics.

Exploring the Future of Engineering: A Conversation with Ole Haaland

Ole Haaland

Biography

Ole Haaland is currently a Robotics Engineer at Prime Vision in Delft, South Holland, where he specializes in areas like test automation, software development, and robotics using tools such as Kubernetes, Git, and Python. He obtained his Master of Science in Cybernetics and Robotics from the Norwegian University of Science and Technology (NTNU). Before joining Prime Vision, Ole worked as an Autopilot Engineer at Tesla in Amsterdam. In this role, he was instrumental in validating Autopilot features and worked closely with the development team. He also contributed as a Robotics Engineer at Maritime Robotics AS in Norway and Aquaai Corporation in Drammen, Viken, Norway, honing his skills in algorithm design, mobile robotics, and software solutions.

Exploring the Future of Engineering: A Conversation with Ole Haaland

By Ole Haaland

Tell us how you're incorporating generative Al into your work?

A couple of years ago I came across this generative Al tool called GitHub Co-Pilot - it's a generative Al system which produces code in the language of your choice. I found it extremely useful, and I didn't want to work without it. I was working in Tesla Autopilot at the time and they had a blanket ban on the tool due to security concerns. But then I started a new job, and luckily they allowed me to use it. But they wouldn't pay for it. I kept insisting to management that this is indeed a tool that everyone should be using. I held a presentation on the topic for my department and convinced quite a few colleagues that this is the future. However, it wasn't until the boom of ChatGPT that my workplace actually started taking these tools seriously. My manager decided that we needed a task force for approaching this in a system making manner. I was approached to take part, due to my strong enthusiasm for the topic. The task force is focused on how we can integrate these tools into the workplace. We're trying to tackle the problems associated with tool integration and trying to understand what people think about them, and what we want to do with them. and how we can use them. The goal is to reach a recommendation of the most helpful and feasible tools But we're also thinking about ethical issues surrounding the use of these tools. Open-sourced solutions like Co-Pilot, or ChatGPT, require us to send all of the data to a server which is outside of our control. So we need to consider ethical questions such as: do we trust these companies with the data? Who owns the code that we generate from these tools?

What is the general perspective of generative Al in the engineering community?

A lot of engineers are quite purist, and rightfully so. They went to university, and dedicated hours of research learning to do things by hand, in an incredibly tedious way. But then, in the real world, you're usually after quick and simple solutions. These Generative Al systems can produce a lot of what engineers have spent hours and hours learning in the blink of an eye. And I think a lot of engineers are skeptical about it - they don't really want to embrace it. So part of the work I'm doing involves, not just finding the right tools, but convincing people that they should

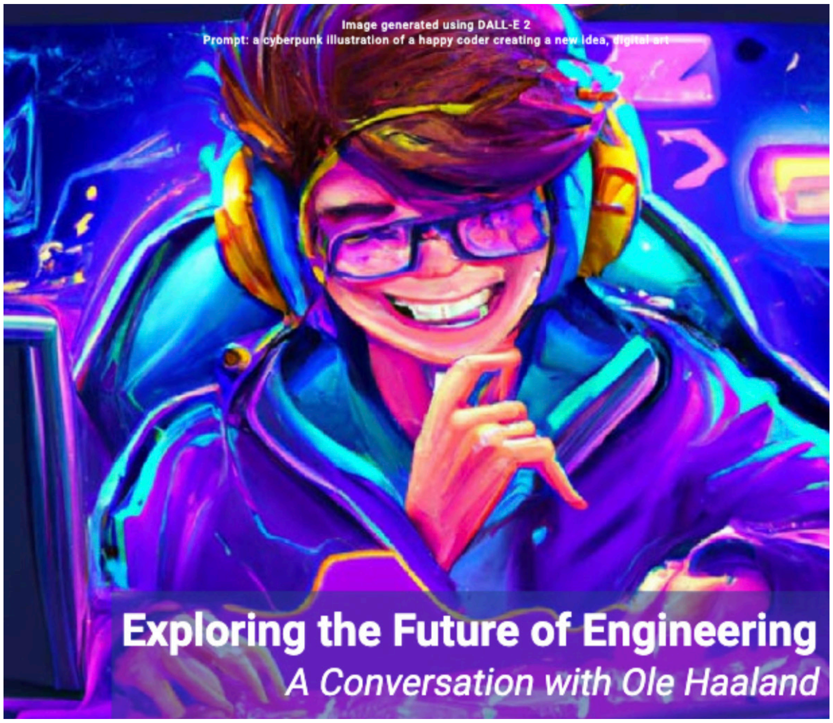


Image generated using DALL-E 2
Prompt: a cyberpunk illustration of a happy coder creating a new idea, digital art
Exploring the Future of Engineering
A Conversation with Ole Haaland

Image generated using DALL-E 2. Prompt: A cyberpunk illustration of a happy coder creating a new idea, digital art use them. But we should keep in mind that these tools aren't perfect. It's not just like you can take these tools and make whatever you want. These tools lack the same understanding that engineers have, so they are prone to making the same error again and again. I think this severely cripples ChatTP's ability to automate your programming job. Without the ability to correctly respond to and resolve errors, there is simply no way for this system to take your job. Someone still needs to understand what is going on. This is why I think these tools will benefit creative people with the ability to understand complex problems. I think the role of the engineer in the future should be seen more as a composer. Or at least more as a composer than an individual musician. While the individual musician focuses on the small details, the composer focuses on how the whole comes together, they're working at the level up. For the past 50 years, computer engineering has been changing constantly. At each stage of development it's become more and more high level, more abstract. Back in the day, we were creating programs by hand, by using punch cards, and then feeding them into the computer. Many of us were still obsessed with single lines of code, asking how we'd format them. But all of that has now gone out the window. The more and more efficient the programming languages, the less work you have to do. From this perspective, ChatGPT is just a case of tool progression. So, yes, ChatGPT is revolutionary, but it can also be seen as another example of us being able to work at a higher level, where yet more details are abstracted away. It might feel revolutionary, but we still require an engineer to understand what's going on. Who knows what it'll be like in 20 years, but I think it'll be fascinating to see.

What does the integration of these tools mean for human collaboration?

A lot of the time you just need to know where to look, and how to look for it. And I guess things have already changed in some way. Way back we used to ask stupid questions to our friends, but now we are addicted to the "google it" mentality. But even with google some people are better at finding results than others. And the same is true for prompting ChatTP in an efficient way. If you don't know the right question, then you won't get the right answer. These tools won't necessarily solve all your problems, you might need help seeing it from a different perspective. While ChatGPT will be a valuable tool for that, I still think human collaboration will be necessary. My personal hope is that these tools will make it easier for us to collaborate, instead of discouraging it. I think and hope that these tools will be used in a way that leaves less admin work for us and more time to discuss what we actually care about. These tools may potentially enhance our communication in the future. For instance, consider an existing email feature that detects and pauses an emotionally charged message before it is sent. Such tools effectively facilitate civil discourse within organizations by promoting better phrasing and tone. Humans can be very rash people, we tend to offend each other and start pointless conflicts. On the other hand, ChatTP is incredibly averse to conflict and would therefore be a great moderator. So in this way, generative Al might actually aid collaboration.

What other positive implications do you see with generative Al models such as ChatGPT?

At its core, I hope these tools will be a very good positive thing, in that they'll allow us to do less work, or more work efficiently. Another thing is that the kind of engineering available to us might be more creative. One incredible thing about ChatGPT is how we can access knowledge without having to dig for stuff. It's a great tool for educating yourself, and this isn't necessarily deep knowledge or understanding, but it helps anyone get their foot in the door for any new topic. Both experts and novices can educate themselves with ChatGPT, and this is a great thing. I'm especially hopeful that these tools will help us manage the information overload that we are exposed to at work. We won't have to check our emails or five different messaging platforms. Instead, information can be condensed and presented to us in an accessible and simple format.

Given your optimistic tone, I'm wondering what you see as the worst possible situation with the

integration of generative Al?

My biggest worry about generative Al boils down to how powerful institutions will make use of them. The most relevant institutions for us in the West are the mega-corporations. Can we really trust these corporations to do the right things with these technologies? At the end of the day, the goal of capitalistic institutions is profit and this is sadly, often not aligned with human flourishing and happiness. A good example of this is how the attention economy has rapidly changed our civilization. Kids are addicted to their phones and I don't think we are too many generations away from a world looking like Wall-E. Generative Al will not exactly slow this trend, but rather help get us even more hooked. What happens when Al generated TikTok becomes the norm? I don't think it will look too good. I also worry about how political institutions could misuse this technology. The ability of ChatGPT to create fake content, or fake conversations is really quite impressive. And, in the wrong hands, this capability of ChatGPT is really concerning. It could be deployed to build trust with people, surveil them or detect the possibility of crimes and things like that. This outcome is something outside of what even George Orwell could imagine. Back in 1984, you just had a camera and a TV. But the applications you can use now to suppress people or control them are really insane.

How has the conversation progressed around generative Al within the engineering community? Is there still the initial fear that systems like ChatGPT are going to take over engineering roles?

I think initially there was a very big hype around it, which is natural. And then there was a sobering period after a couple of weeks, in which people started noticing the flaws in these products. The biggest one is its veracity. Say, you can ask ChatGPT to add 4 and 4 together and it can give you the wrong outcome. Where a 20-year-old calculator would give you the perfect answer. I wouldn't be scared - not right now - there are for sure reasons to be scared, and a lot of traits and skills will become redundant, but I don't think we should fear becoming completely redundant. Yes, ChatGPT can do certain things a hundred times faster than I can. But I don't think we should be scared of this. It just means that we have to rethink what work is and how we do it. And to me, that's freedom. I can make more things in a shorter amount of time. I understand that I have a very optimistic perspective, a lot of people have a much darker take on things. If you are hyper specialized, and not planning on learning something new over the next 20 years, then yes, I'd be incredibly scared about my future career if I were you. But, if you're open to new avenues, and doing new things, then I don't think you need to be scared, you should embrace the change.

What would you want to say to a fellow engineer who has a more pessimistic view about the change that's occurring in your profession right now?

For someone who is just coming into the industry, I'd encourage them to adapt, change, and understand the aspects of their role that will change quickly. However, those of us with long-term memory will remember that this has been said many times before. Embracing change has been a winning trait for many years. And don't give up on trying to understand things deeply. These tools are exactly that - tools - you shouldn't be relying on them to pass your exams and to get by in life. Understanding is key to solving novel problems and that is what engineering is all about. If you lose out on this skill then you will lose in the job market. Use ChatGPT to make yourself smarter, not dumber. For those who are more skeptical. Enjoy doing the boring stuff, the rest of the world will not be waiting around.

KEY TAKEAWAYS

- Our roles, skills, and workplaces are likely to change considerably as generative Al is integrated into our day-to-day. This integration will bring with it many benefits, such as increased productivity and efficiency. However, we will need to take steps to mitigate the potentially negative effects of this transition.

- Artists have throughout history been cautious at the emergence of new tools, but, as with any other technology, generative Al can be used in the right way and the wrong way. These technologies do pose threats to some areas of art, in particular digital art. But, generative Al is just another tool, and giving artists the knowledge and ability to use it in the right way can help to foster rather than hinder artistic creativity.

- Many engineers understandably remain skeptical about generative Al. However, we should bear in mind that these tools will never fully replicate the role of the engineer. The engineer has a level of understanding that an Al system will never have. And engineers will still be needed even as these tools develop. By remaining open to this change, and adapting to the changing landscape, engineers can develop alongside these technologies, rather than against them.

IP Law and Generative AI: Where Are We Now and Where Are We Going?

Larry Sandell

Biography

Larry Sandell is an experienced IP strategist, patent attorney, litigator, and appellate advocate, with a technical focus that includes software—including generative AI and blockchain, consumer electronics, medical devices, cannabis, and food science. Larry has argued in the U.S. Courts of Appeal for the Federal Circuit, the Ninth Circuit, and the D.C. Circuit, and has litigated before numerous Federal District Court, the Patent Trial and Appeal Board (PTAB), the Trademark Trial and Appeal Board (TTAB), and the US International Trade Commission (ITC). As IP law catches up to the AI age, Larry is applying his skills and expertise to helping clients not only manage AI-related risk, but position themselves to thrive in this brave new world. Eager to experiment with new AI tech applications, Larry's hobbies include using ChatGPT and DALL-E to write and illustrate love poems to his wife and stories for his young children. Larry practices law with Mei & Mark LLP and can be reached at lsandell@meimark.com.

IP Law and Generative AI: Where Are We Now and Where Are We Going?

By Larry Sandell

As to generative AI, Intellectual Property (IP) law is murky and far from settled, but some answers, legal strategies, and best practices can be derived from existing law. Copyright law in particular is already struggling with how to fairly allocate creator attribution and ownership of AI outputs, whether and to what degree AI outputs infringe upon underlying works, and even what liability may attach to the use of copyrighted dataset inputs. While the questions on this topic are legion, this article provides an overview of the current state of US IP law as it applies to generative AI and some insight as to where it is headed.

Who owns the content created by generative AI?

Unsurprisingly, the answer here is very fact-dependent—but also without clear, established legal parameters. Because IP rights can generally be assigned by contract, a good place to start is with the user terms governing a generative AI system. As a prominent example, OpenAI purports to assign ownership of text and art outputs to the user only if the user complies with its terms of service. However, OpenAI's terms prohibit the use of generative AI "in a way that infringes, misappropriates or violates any person's rights." Here, OpenAI appears to pass the buck on IP infringement liability, which, as discussed below, cannot be readily or reliably assessed. As a result, a user cannot be assured of her ownership of OpenAI's output.

Are generative AI outputs protectable by copyright?

To be eligible for copyright protection under US law, text or images must be "original works of authorship fixed in a tangible medium of expression." But, what does this mean content generated entirely or partially by AI?

The US Copyright Office has determined that generative AI cannot be an "author" under copyright law. It unequivocally rejected the attempted registration of a digital painting where a generative AI platform was declared to be the work's "author." The registration request was denied because the work lacked "human authorship," and this rejection was recently affirmed by the Federal District Court in DC. The Federal Judge took guidance from the "monkey selfie" case, wherein the Ninth Circuit Court of Appeals held that the Copyright Act did not permit a crested macaque (represented by PETA) to sue for infringement regarding photos the primate took on a nature photographer's camera. Ultimately, the court confirmed that non-humans cannot be "authors" under US copyright law.

A more practical question on this topic is whether a human can be an author when generative AI is heavily used as a creative tool. The issue was squarely raised by Zarya of the Dawn, a graphic novel prepared with substantial generative AI assistance. Zarya's human creator filed for copyright registration on a comic book made with MidJourney-generated images. While the Copyright Office initially registered copyrights on both the images and the overall compilation of the book, it ultimately canceled the copyright on the AI-generated images. The arguments that AI's contribution was merely "assisting";

that the book reflected the human author's creative, iterative AI prompts and overall artistic vision; and that the AI outputs were manually modified before publication were all rejected by the Office.

Until and unless the Copyright Office's position on AI-generated content is appealed and overturned by a Federal Court, it is the law. Nonetheless, it seems reasonably likely generative AI may ultimately be accepted as an "assisting" tool that does not preclude copyright—at least in some circumstances. Indeed, the prospect of copyright protection for works partly generated by AI has become relevant to the ongoing strike by the Writers Guild of America against the Alliance of Motion Picture and Television Producers.

The advent of photography raised a similar debate: Some argued that photographers merely engaged in a rote process without creative input. But, in the late 1800s, Congress expressly made photographs copyrightable. Since then, the Supreme Court has held that copyright merely requires "at least some minimal degree of creativity" "no matter how crude, humble or obvious" the "creative spark" might

Stephen Thaler's AI creation, A Recent Entrance to Paradise, has been denied copyright protection by the US Copyright Office.

Are generative AI outputs protectable by patent law?

This question has been answered with clarity—at least for now. In August 2022, the Federal Circuit Court of Appeals affirmed "'inventors' must be human." The more practical issue, however, concerns whether generative AI's contribution to technological invention can undermine patentability. Under US patent law, inventorship is split into two parts—(1) "conception" of the idea underlying the claimed invention and (2)"reduction to practice," that is, bring the idea to technical fruition. Only persons involved in "conception" are "inventors"; those that merely reduce the invention to practice—regardless of the technical skill and effort contributed—are not. While no case has addressed what level of AI contribution might preclude patentability, human inventors should be able to patent their inventions at least when a generative AI contribution is limited to "reduction to practice."

Where is the line between "fair use" and copyright infringement via creation of derivative works?

Generative AI training datasets commonly include copyrighted works—including art, photos, writings, sound recordings, video, and code. These inputs are inherently used to generate AI output. Yet, whether and to what degree these uses of copyrighted works are permissible are open questions.

On one hand, a copyright holder possesses the exclusive right to create derivative works—namely, creative works based on their original (copyrighted) work. AI synthesizes and utilizes images, literature, and code when it generates output—but does this mean that generative AI output is a derivative work? If so, each AI platform (and its owner) may infringe copyrights on a regular basis. On the other hand, the "fair use" doctrine can immunize certain uses of copyrighted works based on, for example, (1) the commercial vs non-profit/educational nature of the use and how transformative the use is; (2) if the original works are more factual/technical vs. artistic/creative (because copyright protects expression, not underlying data); (3) the amount of the original works used and how identifiable the elements may be; and (4) the effect on the commercial value of the original work.

In late September 2023, the U.S. District Court in Delaware became the first court to weigh in. In that case, Thomas Reuters, owner of the Westlaw legal services platform, had accused Ross Intelligence, a legal-research industry upstart, of copyright infringement regarding its AI system that permits a user to input a question and receive an responsive quote from a legal opinion as an answer. Thomson Reuters had taken exception to Ross's use of Westlaw's copyrighted "headnotes" (categorized summaries of legal opinions) paired with corresponding quotes from legal opinions as AI platform inputs. The Court declined to find for either party on summary judgment on the fair use question because many factual disputes remained, effectively punting it to the jury in a future trial. Notably, the Court explained that the nature of AI outputs of Ross' platform—including an assessment of their "transformativeness" vis-à-vis the original Westlaw works—would be critical to a final fair use determination.

But outside of this recent reminder that the fair use inquiry is highly fact-dependent, U.S. Courts have, to date, offered practically no insight as to where the line between fair use and infringement via derivative works might be in various generative AI contexts. There are, however, several important cases to watch on this issue:

First, in November 2022, GitHub, Microsoft, and OpenAI were sued by anonymous coders who contributed to Github's open source code repository under the open source GNU General Public License. The coders argue that training the for-profit Codex and CoPilot generative AI platforms on Github's open source code repository both breaches the contractual terms of the GNU license and impermissibly removes copyright management text (e.g., human-readable references to the GNU license in each section of code). The companies argued that the case should be dismissed because there is no evidence of actual violations in AI output—merely assumptions of wrongdoing based on the training dataset. In May 2023, the U.S. District Court for the Northern District of California dismissed most of the Plaintiffs' claims for failing to allege sufficient underlying facts. The coders have since amended their complaint to flesh out their allegations, and the software firms have responded by again seeking an early end to the case.

Second, in another class action suit, Stability AI, MidJourney, and Deviant Art were sued in January 2023 by a group of digital visual artists. The artists assert copyright, publicity, and unfair competition claims, arguing that "AI image generators are 21st-century collage tools that violate the rights of millions of artists" and, in particular, that that the AI generation of commission-free art made "in the style of" a particular artist erodes their commercial opportunities. In late July, the Court held a hearing regarding the AI firms' motions to dismiss the case; the Court is expected to issue a written ruling shortly.

Third, in February 2023, Getty Images filed suit against Stability AI asserting copyright, trademark and unfair competition claims. Getty asserts "Stability AI has copied more than 12 million photographs from Getty Images' collection, along with the associated captions and metadata, without permission from or compensation to Getty Images, as part of its efforts to build a competing business." Getty identifies Stable Diffusion's output images that substantially copy original works, remove or alter copyright notices, and mangle the Getty trademark. Stability has sought to dismiss the case, largely arguing that its UK entity cannot be sued in Delaware, where the case was filed.

More recently, in July 2023, comedian Sarah Silverman and other authors filed sister class action lawsuits against OpenAI and Meta in the Northern District of California. The authors allege copyright infringement, unfair competition, and improper removal of copyright management information because of the tech firms' alleged utilization of their copyrighted text in training large language models. These cases will begin to shape US law on generative AI and copyright. Ultimately, however, Congress may consider resolving some of these issues by mandating a compulsory licensing that requires generative AI platforms to identify copyrighted works materially utilized in a particular AI output (something that might be difficult given that sources of output are often unidentifiable in current AI models)—and attribute and compensate the copyright holders accordingly. An analogous US law has long provided for compulsory licensing payments to compensate songwriters for others'; recordings of their covers.

What can Businesses and Content Creators do to Protect Themselves in this Uncertain and Shifting Legal Landscape?

Although the Copyright Office is seeking formal public comments on key AI-copyright policy issues (until October 18, 2023), clarity in the law will likely continue to lag well behind AI innovation for the foreseeable future. As a practical matter, businesses and artists simply cannot wait manage risk and protect their livelihoods.

So far, enterprise businesses have largely avoided AI-generated content due to substantial copyright infringement liability risk. However, at least one major player has sought to become an early market leader by leveraging this legal uncertainty. Specifically, Adobe's Firefly image platform provides infringement indemnity provisions to assure its customers. Adobe can afford this approach—as opposed to OpenAI's "pass the buck" liability strategy—because it limits Firefly's training set to materials it holds copyright in and public domain works. Businesses desiring AI-generated content should hew closely to platforms that offer reps and warranties that meaningfully mitigate copyright liability risk—even if they need to pay a premium for it.

Artists lacking the resources or the desire to engage in court battles may find that their best recourse is to fight technology with technology. An early leader here is Glaze, a system that alters digital images in a way that is minimally perceptive to the human eye, but hinders the ability of AI diffusion models to process the images in a meaningful way. By "glazing" their work before distribution, a digital visual artist may hamper generative AI platforms' ability to mimic her artistic style.

Additionally, the prospect of widespread generation of infringing derivative works by AI, has made it even more important to ensure the potential availability for statutory copyright damages. Statutory damages enable a copyright holder to receive monetary damages in a lawsuit without specifically proving the commercial value of damage caused by the infringement—a task for which evidence can be difficult or impossible to find. Eligibility for statutory damages requires copyright registration within three months of publication or infringement of a work. As a result, content creators of all stripes should consider filing with the US copyright office every three months to maximize enforcement possibilities in the future. Such filings are surprisingly low cost, and the Copyright Office even permits the registration of computer code in a manner that maintains trade secret status for the code.

Companies with generative AI platforms—and those that publish generative AI output—should also consider prominently offering the public the opportunity to directly lodge complaints about infringement. As a practical matter, this may stave off some lawsuits or demonstrate good faith in court. For example, OpenAI's terms of service wisely include a procedure for receiving copyright complaints, potentially taking advantage of the copyright liability safe harbor provisions of the Digital Millennium Copyright Act.

Conclusion

Inevitably, the evolution of IP law on generative AI will continue to lag behind the rapid evolution of the technology itself. This poses strategic problems for generative AI firms and their clients because IP law—and copyright law in particular—may ultimately render certain products and services commercially unviable due to legal liability. For now, however, the best course of action appears to be: innovate, stay informed, be fair, and hedge legal risks.

KEY TAKEAWAYS

- When it concerns generative AI, Intellectual Property (IP) law is murky and far from settled, but some answers, legal strategies, and best practices can be derived from existing law.

- Copyright law in particular is already struggling with how to fairly allocate creator attribution and ownership of AI outputs, whether and to what degree AI outputs infringe upon underlying work, and even what liability may attach to the use of copyrighted dataset inputs.

- While the questions on this topic are legion, this article seeks to provide an overview of the current state of US IP law as it applies to generative AI and some insight as to where it is headed.

- Unsurprisingly, the answer here is very fact-dependent—but also without established legal parameters. A good place to start is with the user terms governing a generative AI system.

- As a prominent example, OpenAI purports to assign ownership of text and art outputs to the user only if the user complies with its terms of service. However, OpenAI's terms prohibit the use of generative AI "in a way that infringes, misappropriates or violates any person's rights".

How to Apply Generative AI to Business: An Exploration of Use Cases

Grid Dynamics | Contributors

Ilya Katsov

VP of Technology

Rohit Tripathi

Principal, CTO Office

Eugene Steinberg

Technical Fellow, Head of Digital Commerce

Sethuram Sankarasubramaniam

Director Customer Success -Life Science & Pharma

Leo Shulman

VP, Insurance Practice Lead

How to Apply Generative AI to Business: An Exploration of Use Cases

By Grid Dynamics

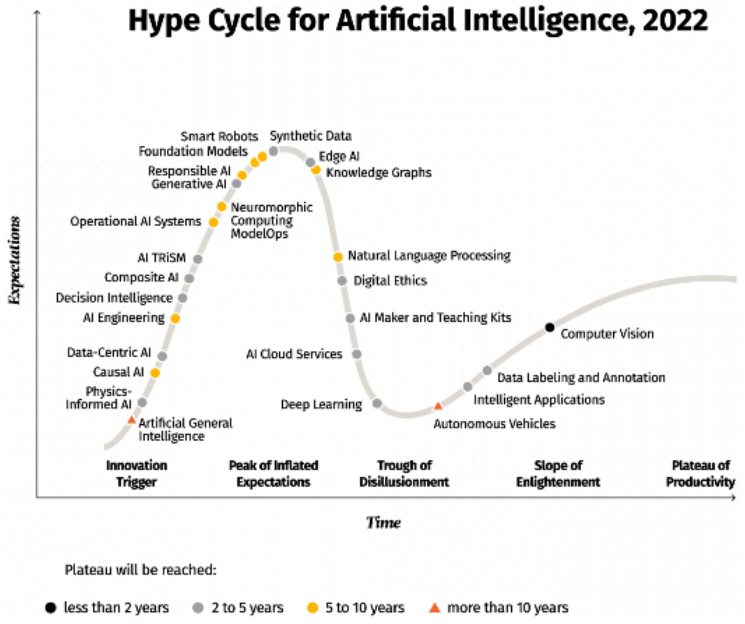
Introduction

After 25+ years of innovation, the field of artificial intelligence (AI) has entered a new stage due to the breakthroughs in generative AI technology. Goldman Sachs projects that generative AI could raise the global GDP a full 7% within 10 years.

Generative AI uses past data to produce entirely new content, such as text, images, videos, computer code, or synthetic data, drawing inspiration from its training. As industries look for ways to leverage this technology to gain a competitive edge, business executives wonder: "How can I apply generative AI to my business?"

While generative AI will not replace humans, businesses shouldn't attempt to automate human jobs but rather to leverage AI to assist their existing human workforce, increasing the efficiency of their employees and ensuring process optimization.

In this article, we will explore generative AI applications across industries and address the following topics:
1. Necessary foundations for generative AI applications
2. Production-ready use cases
3. Mid- and long-term applications, and
4. The risks of generative AI.



Foundational ecosystem for generative AI — data, processes, IoT and cloud

According to Gartner's AI hype cycle, generative AI is nearing its peak of expectations, with many industry leaders still unclear on its fundamentals of the technology. The media hype and lack of clear strategies by companies are leading to a bandwagon effect. As a result, overhyped investments may soon give way to disappointments.

That being said, generative AI is poised to enhance the productivity of enterprise workforces and boost the efficiency of human tasks. To lay the foundation for this transformation, directors need to build an ecosystem consisting of three robust business capabilities that allow generative AI to thrive:

1. Get your data quality in order

Corrupted data can lead to imprecise AI decisions, reducing the accuracy and confidence of insights. Adding data quality monitoring to the data lake is essential to prevent defects.

2. Leverage the cloud for efficient storage and computing power

AI-driven innovations often lead to cloud adoption as the cloud's scalability and ease of infrastructure setup cut costs and enhance DataOps and MLOps quality. Traditional on-premise data systems are simply outdated, costly, and lack the scalability requirements demanded by

3. Enable modern ecosystems to collect IoT data

Businesses often have IoT devices that interact and share data. To leverage these devices, it's of paramount importance that businesses consolidate and manage data across channels. Transitioning away from outdated systems can enhance data visualization and allow for AI experimentation in a modern and responsive ecosystem, optimized for IoT.

ENTERPRISE USE CASES: CRAWL, WALK, RUN

Crawl: Start with a low-risk production-ready use case

Walk: Experiment with various tools to solve problems innovatively

Run: Get your entire organization on board to maximize output

Healthcare, Pharma, and Medtech

Healthcare Providers (HCPs) spend an average of 26.7 hours daily on administrative duties and patient care, with two-thirds of their work being non-patient-facing. Generative AI can significantly benefit healthcare by easing data access, reducing physician burnout, and automating tasks.

Applications of generative AI in Pharma

Crawl: Start with a low-risk production-ready use case

Ready-to-use resources for HCPs: Pharma companies can use LLM-based AI to transform

Accelerated KOL video production: Generative AI tools can reduce video production time and costs of filming Key opinion leaders (KOLs) who are external experts aiding pharma companies.

Walk: Experiment with various tools to solve problems innovatively

Empowered sales teams ensure successful HCP engagement: Generative AI can analyze HCP data to enhance customer journeys, tailor messaging, and avoid oversaturating HCPs

Run: Get your entire organization on board to maximize output

Accelerated drug development: Generative AI can analyze clinical trial data to identify drug targets and predict effective compounds, speeding up drug development and reducing costs.

Digital clinical trials: Digital trials improve participant enrollment, engagement, and trial quality as wearable tech data can monitor participants. Generative AI can simplify trial results, making clinical treatments more globally accessible.

Predictive maintenance: Generative AI can analyze medical device data to predict maintenance needs, enabling HCPs to proactively manage equipment and minimize failure risks.

Applications of generative AI in Healthcare

Crawl: Start with a low-risk production-ready use case

Assisted clinical documentations: LLM-based generative AI that automates patient summaries frees up HCPs for vital tasks.

Writing referrals: LLMs can auto-generate referral letters for specialists using EHRs, freeing up time and quickening diagnosis.

Diversifying healthcare data: With prompt-based guardrails, tools can generate synthetic healthcare data to address diversity gaps in medical imaging.

Walk: Experiment with various tools to solve problems innovatively

Automated clinical note-taking: Using voice-to-text AI models, clinical documentation can be automated based on HCP-patient conversations, making note-taking easier.

Image-to-text clinical summaries: Image-to-text models can offer visual diagnosis summaries for HCPs to verify.

Run: Get your entire organization on board to maximize output

Personalized medicine: Generative AI can analyze vast data, like health drivers and genomic information, to enhance care by identifying patterns and predicting outcomes, allowing for personalized treatments.

Medical image resolution: New GAN-based architectures can convert low-resolution medical images (MRIs, X-rays, CT) to high-resolution.

Manufacturing

Manufacturing leaders are moving towards Industry 4.0, aiming for smart factories with autonomous decision-making systems. They can use generative AI for critical thinking and decision-making, like creating digital prototypes.

Applications of generative AI in Manufacturing

Crawl: Start with a low-risk production-ready use case

Product design ideation and optimization: Generative AI can generate designs from textual descriptions, optimize existing designs, and create digital twins.

Streamlined employee training, enhanced procurement documentation, and ticket resolution: LLMs can offer on-demand training for manufacturing aspects like safety and quality control. They can also assist procurement teams in drafting sourcing plans from past data, and replace internal support with conversational NLP interfaces.

Walk: Experiment with various tools to solve problems innovatively

Making dashboards more human: By integrating supply chain dashboards with LLM-based solutions, manufacturers can simplify complex data for all, enabling quicker resolution of quality issues.

Run: Get your entire organization on board to maximize output

Infrastructure process optimization and quality control using digital twins: Using data from the supply chain and generative AI, manufacturers can create digital twins to detect defects during product development, often overlooked by humans.

Material discovery: Modern generative design models incorporate more constraints than just volume for old topology optimization algorithms, allowing AI to evaluate various materials and manufacturing principles, streamlining engineering.

Financial services and insurance

The financial services industry has been progressively focusing on enhancing customer experiences using technology in recent years. Generative AI's potential is vast but comes with regulatory challenges that need careful navigation to ensure compliance and data protection.

Applications of generative AI in Financial Services

Crawl: Start with a low-risk production-ready use case

Drafting technical requirements for new products: Generative AI can transform transcripts of online meetings including technical specifications into technical requirements, streamlining project planning.

Walk: Experiment with various tools to solve problems innovatively

Virtual financial advisor: Generative AI can analyze vast data, offering financial insights. Financial institutions can use AI-driven apps to guide customers based on their finances.

Attributing customers to correct segments: Models can categorize customers based on behavior, offering flexibility and cost savings which aids banks in product targeting.

Understanding customer sentiments and exploring upselling and cross-selling opportunities: By analyzing customer-support interactions, generative AI can gauge customer sentiment, aiding in timely upselling and cross-selling.

Run: Get your entire organization on board to maximize output

"Run" applications in banking involve generative AI for daily operations with sensitive customer data. Currently, generative AI lacks adequate regulation for these tasks due to the industry's susceptibility to fraud.

Applications of generative AI in Insurance

Crawl: Start with a low-risk production-ready use case

Social enablement of insurance agents using AI-generated educational resources: As insurance becomes more digital, agents grapple with new tech for lead generation and must draft detailed cover letters for potential customers. LLMs can simplify these tasks.

Creating quick FAQs for every policy: Use LLMs to transform complex policy information into user-friendly FAQs, enhancing the customer service experience.

Walk: Experiment with various tools to solve problems innovatively

Generative AI virtual insurance advisors for product discovery: To stand out in a competitive market, insurance companies should utilize a generative AI chatbot for product discovery and related FAQs.

Run: Get your entire organization on board to maximize output

Speeding up the underwriting process: Administrative tasks leading to a $160 billion efficiency loss in five years for underwriters. Generative AI can assess applicant risk profiles and highlight data gaps, streamlining underwriters' tasks.

Fraud defense in claim management: Fraud costs financial institutions billions annually. Generative AI can analyze data like transaction histories and credit scores to detect and prevent fraud.

Risk assessment: Financial institutions require precise market trend predictions for informed decisions. Generative AI can analyze market data and customer behavior to pinpoint risks and opportunities.

Gaming

AI tools are revolutionizing the gaming industry by speeding up game development, customizing content, and reducing costs. But they still require human creativity for truly innovative game design, as they struggle with generating accurate content without sufficient existing information.

Crawl: Start with a low-risk production-ready use case

Creating concept art: Generative AI aids in ideation, allowing designers to generate images with tools, significantly cutting image production time.

Generate game development steps using LLMs and modify the rules of existing games: LLMs can guide new game developers through game creation, or modify existing games with your rules to craft a new one.

Walk: Experiment with various tools to solve problems innovatively

Creating 2D and 3D content, rapid prototyping, and more: The gaming industry invests $60 billion annually in content creation. Generative AI tools expedite the creation of 2D and 3D assets, characters, and settings, speeding up prototyping and experimentation.

Run: Get your entire organization on board to maximize output

AI-generated NPCs: AI-generated NPCs enhance game immersion with realistic and adaptive content, adding depth and dynamic interactions to the gameplay.

Speech, dialog, and music generation: Companies are developing realistic NPC voices using generative AI, moving away from pre-recorded voice actor dialogues. AI also allows adaptive music that aligns with on-screen events.

Retail

Retailers use first-party data and personalization campaigns to enhance customer loyalty and reduce churn, but with the digital noise, generative AI emerges as a solution: acting like a proactive digital store assistant, optimizing messaging and enhancing the shopping experience in real-time.

Applications of generative AI in Retail

Crawl: Start with a low-risk production-ready use case

Compelling product descriptions for better SEO and subtle personalization: Generative AI helps retailers standardize and optimize product titles and can purposely align with a brand's tone, allowing personalized descriptions.

Improving product attribution: Can use LLMs to analyze and enrich product data, enhancing data fidelity and customer experience.

Personalization using streamlined 2D and 3D product modeling: Generative AI can generate photo-realistic product images from textual prompts, eliminating the need for physical production and allowing customers to see diverse product representations.

Walk: Experiment with various tools to solve problems innovatively

Showcasing products in different environments and according to themes: Using AI tools, retailers can showcase products in diverse backgrounds, helping customers visualize the product's versatility.

Conversational AI for enhanced customer service: Generative AI can offer intelligent shopping suggestions based on a user's search history and streamline contact center operations, increasing consumer satisfaction.

Conversational product discovery and selection: Generative AI can further enhance search capabilities by acting as a virtual shopping assistant, understanding language nuances and analyzing product details for a personalized experience.

Run: Get your entire organization on board to maximize output

Virtual models for inclusive fashion: Generative AI can generate personalized outfit visuals with virtual models, offering cost savings and flexibility in appearance, lighting, and poses.

Virtual try-ons and improved transaction flows: Generative AI can enhance current personalizations by creating personalized visuals and customized web pages, offering unique site experiences based on user preferences and behaviors.

RISKS IN LEVERAGING GENERATIVE AI TODAY

1. Market disruption

Google's dominance in search was challenged by ChatGPT in 2022, highlighting generative AI's disruptive potential across various industries, not just low-skilled jobs. Many businesses, fresh from digital transformations, now face another shift due to AI. A Goldman Sachs report suggests generative AI might eliminate 300 million jobs globally, including 19% in the U.S. This technological shift, unlike past disruptions, may not benefit everyone, with many companies undergoing significant changes soon.

2. Reputation at stake

Industries value racial and gender diversity. However, many AI solutions, trained on biased data, have caused PR issues. Microsoft's Bing made errors during its demo, highlighting that AI can't replace human judgment. It's crucial to use generative AI cautiously, understanding its potential risks to an organization's reputation before deployment.

3. Explosion of cybercrimes

Due to the rise of scammers and hackers, businesses have become more aware of data and system security. Data breaches cost millions, and generative AI could amplify these threats. As AI advances, managing cyber risks will become more challenging, surpassing what traditional firewalls can handle.

4. Legal implications

The European Commission has drafted the AI Act to regulate booming generative AI technologies. This law requires companies to disclose copyrighted materials used in AI development. As global governments introduce similar legislation, companies must consider not just compliance but also liability issues. If an AI-driven product fails, the organization, AI developer, or data provider could be held responsible. It's crucial to maintain transparency about AI decision-making processes.

5. Data Exposure

Implementing advanced technology like generative AI can offer benefits, but it's not without risks. Mistakes can lead to loss of trade secrets or financial setbacks. Even top AI systems, like ChatGPT, can have flaws; for instance, a bug exposed payment details of some users in 2023. It's crucial for business leaders to weigh these operational risks against the allure of innovation.

THE NEXT STEP: CONSUME OR CUSTOMIZE

Generative AI and LLMs are like tools in a toolbox: available but generic. Using them directly may pose risks due to limited effectiveness. Customizing these tools, akin to tailoring them for specific tasks, ensures they're optimized for your business. A well-structured digital ecosystem lets AI access the right data, enhancing organizational readiness. Customized AI models, while requiring more effort, offer better results, control, and reduced risks.

Generative AI is not a fleeting technological trend; it's a force of transformation. Its potential to reshape industries is vast, but the key to success lies in understanding its capabilities, laying a robust foundation, and integrating it strategically into business operations. As businesses navigate this new frontier, the path forward is illuminated with promise and potential, but it demands clarity, strategy, and foresight.

KEY TAKEAWAYS

- **Generative AI's Potential and Impact:** Generative AI, which creates new content based on past data, is projected to increase global GDP by 7% within a decade. While it won't replace humans, it can significantly enhance workforce productivity and process optimization across various industries.

- **Foundational Ecosystem for Generative AI:** For generative AI to be effective, businesses need to ensure data quality, leverage cloud computing for storage and processing, and integrate modern ecosystems to collect IoT data.

- **Industry-specific Applications of Generative AI:**
  - Healthcare: Generative AI can reduce administrative tasks, accelerate drug development, and enhance patient care.
  - Manufacturing: AI can aid in product design, streamline training, and optimize infrastructure processes.
  - Financial Services: Generative AI can streamline technical requirements, offer virtual financial advice, and enhance customer segmentation.
  - Gaming: AI tools can expedite game development, enhance game immersion, and create realistic NPC interactions.
  - Retail: Generative AI can optimize product descriptions, enhance customer service, and provide personalized shopping experiences.

- **Risks of Generative AI:** The adoption of generative AI presents challenges such as potential market disruption, reputational risks, increased cybercrimes, legal implications, and data exposure risks.

- **Customization vs. Direct Consumption:** While generative AI tools are available, using them directly might not be effective for specific business needs. Customizing these tools ensures better results, control, and reduced risks. The success of generative AI integration hinges on understanding its capabilities and strategically incorporating it into business operations.

Managing the Risks of Generative AI

Kathy Baxter

Biography

Kathy Baxter is Principal Architect of Ethical AI Practice at Salesforce, developing research-informed best practices to educate Salesforce employees, customers, and the industry on the development of responsible AI. She collaborates and partners with external AI and ethics experts to continuously evolve Salesforce policies, practices, and products. She is a member of Singapore's Advisory Council on the Ethical Use of AI and Data, Visiting AI Fellow at NIST, and on the Board of EqualAI. Prior to Salesforce, she worked at Google, eBay, and Oracle in User Experience Research. She is the co-author of "Understanding Your Users: A Practical Guide to User Research Methodologies."

Managing the Risks of Generative AI

Yoav Schlesinger

Biography

Yoav Schlesinger is an Architect of Ethical AI Practice at Salesforce, helping the company embed and instantiate ethical product practices to maximize the societal benefits of AI. Prior to coming to Salesforce, Yoav was a founding member of the Tech and Society Solutions Lab at Omidyar Network, where he launched the Responsible Computer Science Challenge and helped develop EthicalOS, a risk mitigation toolkit for product managers.

Managing the Risks of Generative AI

By Kathy Baxter and Yoav Schlesinger

Corporate leaders, academics, policymakers, and countless others are looking for ways to harness generative AI technology, which has the potential to transform the way we learn, work, and more. In business, generative AI has the potential to transform the way companies interact with customers and drive business growth. New research shows 67% of senior IT leaders are prioritizing generative AI for their business within the next 18 months, with one-third (33%) naming it as a top priority. Companies are exploring how it could impact every part of the business, including sales, customer service, marketing, commerce, IT, legal, HR, and others.

However, senior IT leaders need a trusted, data-secure way for their employees to use these technologies. Seventy-nine-percent of senior IT leaders reported concerns that these technologies bring the potential for security risks, and another 73% are concerned about biased outcomes. More broadly, organizations must recognize the need to ensure the ethical, transparent, and responsible use of these technologies.

A business using generative AI technology in an enterprise setting is different from consumers using it for private, individual use. Businesses need to adhere to regulations relevant to their respective industries (think: healthcare), and there's a minefield of legal, financial, and ethical implications if the content generated is inaccurate, inaccessible, or offensive. For example, the risk of harm when an generative AI chatbot gives incorrect steps for cooking a recipe is much lower than when giving a field service worker instructions for repairing a piece of heavy machinery. If not designed and deployed with clear ethical guidelines, generative AI can have unintended consequences and potentially cause real harm.

Organizations need a clear and actionable framework for how to use generative AI and to align their generative AI goals with their businesses' "jobs to be done," including how generative AI will impact sales, marketing, commerce, service, and IT jobs.

In 2019, we published our trusted AI principles (transparency, fairness, responsibility, accountability, and reliability), meant to guide the development of ethical AI tools. These can apply to any organization investing in AI. But these principles only go so far if organizations lack an ethical AI practice to operationalize them into the development and adoption of AI technology. A mature ethical AI practice operationalizes its principles or values through responsible product development and deployment — uniting disciplines such as product management, data science, engineering, privacy, legal, user research, design, and accessibility — to mitigate the potential harms and maximize the social benefits of AI. There are models for how organizations can start, mature, and expand these practices, which provide clear roadmaps for how to build the infrastructure for ethical AI development.

But with the mainstream emergence — and accessibility — of generative AI, we recognized that organizations needed guidelines specific to the risks this specific technology presents. These guidelines don't replace our principles, but instead act as a North Star for how they can be operationalized and put into practice as businesses develop products and services that use this new technology.

Guidelines for the Ethical Development of Generative AI

Our new set of guidelines can help organizations evaluate generative AI's risks and considerations as these tools gain mainstream adoption. They cover five focus areas.

Accuracy

Organizations need to be able to train AI models on their own data to deliver verifiable results that balance accuracy, precision, and recall (the model's ability to correctly identify positive cases within a given dataset). It's important to communicate when there is uncertainty regarding generative AI responses and enable people to validate them. This can be done by citing the sources where the model is pulling information from in order to create content, explaining why the AI gave the response it did, highlighting uncertainty, and creating guardrails preventing some tasks from being fully automated.

Safety

Making every effort to mitigate bias, toxicity, and harmful outputs by conducting bias, explainability, and robustness assessments is always a priority in AI. Organizations must protect the privacy of any personally identifying information present in the data used for training to prevent potential harm. Further, security assessments can help organizations identify vulnerabilities that may be exploited by bad actors (e.g., "do anything now" prompt injection attacks that have been used to override ChatGPT's guardrails).

Honesty

When collecting data to train and evaluate our models, respect data provenance and ensure there is consent to use that data. This can be done by leveraging open-source and user-provided data. And, when autonomously delivering outputs, it's a necessity to be transparent that an AI has created the content. This can be done through watermarks on the content or through in-app messaging.

Empowerment

While there are some cases where it is best to fully automate processes, AI should more often play a supporting role. Today, generative AI is a great assistant. In industries where building trust is a top priority, such as in finance or healthcare, it's important that humans be involved in decision-making

— with the help of data-driven insights that an AI model may provide — to build trust and maintain transparency. Additionally, ensure the model's outputs are accessible to all (e.g., generate ALT text to accompany images, text output is accessible to a screen reader). And of course, one must treat content contributors, creators, and data labelers with respect (e.g., fair wages, consent to use their work).

Sustainability

Language models are described as "large" based on the number of values or parameters it uses. Some of these large language models (LLMs) have hundreds of billions of parameters and use a lot of energy and water to train them. For example, GPT3 took 1.287 gigawatt hours or about as much electricity to power 120 U.S. homes for a year, and 700,000 liters of clean freshwater.

When considering AI models, larger doesn't always mean better. As we develop our own models, we will strive to minimize the size of our models while maximizing accuracy by training on models on large amounts of high-quality CRM data. This will help reduce the carbon footprint because less computation is required, which means less energy consumption from data centers and carbon emission.

Integrating Generative AI

Most organizations will integrate generative AI tools rather than build their own. Here are some tactical tips for safely integrating generative AI in business applications to drive business results:

Use zero-party or first-party data

Companies should train generative AI tools using zero-party data — data that customers share proactively — and first-party data, which they collect directly. Strong data provenance is key to ensuring models are accurate, original, and trusted. Relying on third-party data, or information obtained from external sources, to train AI tools makes it difficult to ensure that output is accurate.

For example, data brokers may have old data, incorrectly combine data from devices or accounts that don't belong to the same person, and/or make inaccurate inferences based on the data. This applies for our customers when we are grounding the models in their data. So in Marketing Cloud, if the data in a customer's CRM all came from data brokers, the personalization may be wrong.

Keep data fresh and well-labeled

AI is only as good as the data it's trained on. Models that generate responses to customer support queries will produce inaccurate or out-of-date results if the content it is grounded in is old, incomplete, and inaccurate. This can lead to hallucinations, in which a tool confidently asserts that a falsehood is real. Training data that contains bias will result in tools that propagate bias.

Companies must review all datasets and documents that will be used to train models, and remove biased, toxic, and false elements. This process of curation is key to principles of safety and accuracy.

Ensure there's a human in the loop

Just because something can be automated doesn't mean it should be. Generative AI tools aren't always capable of understanding emotional or business context, or knowing when they're wrong or damaging. Humans need to be involved to review outputs for accuracy, suss out bias, and ensure models are operating as intended. More broadly, generative AI should be seen as a way to augment human capabilities and empower communities, not replace or displace them.

Companies play a critical role in responsibly adopting generative AI, and integrating these tools in ways that enhance, not diminish, the working experience of their employees, and their customers. This comes back to ensuring the responsible use of AI in maintaining accuracy, safety, honesty, empowerment, and sustainability, mitigating risks, and eliminating biased outcomes. And, the commitment should extend beyond immediate corporate interests, encompassing broader societal responsibilities and ethical AI practices.

Test, test, test

Generative AI cannot operate on a set-it-and-forget-it basis — the tools need constant oversight. Companies can start by looking for ways to automate the review process by collecting metadata on AI systems and developing standard mitigations for specific risks.

Ultimately, humans also need to be involved in checking output for accuracy, bias and hallucinations. Companies can consider investing in ethical AI training for front-line engineers and managers so they're prepared to assess AI tools. If resources are constrained, they can prioritize testing models that have the most potential to cause harm.

Get feedback

Listening to employees, trusted advisors, and impacted communities is key to identifying risks and course-correcting. Companies can create a variety of pathways for employees to report concerns, such as an anonymous hotline, a mailing list, a dedicated Slack or social media channel or focus groups. Creating incentives for employees to report issues can also be effective.

Some organizations have formed ethics advisory councils — composed of employees from across the company, external experts, or a mix of both — to weigh in on AI development. Finally, having open lines of communication with community stakeholders is key to avoiding unintended consequences.

With generative AI going mainstream, enterprises have the responsibility to ensure that they're using this technology ethically and mitigating potential harm. By committing to guidelines and having guardrails in advance, companies can ensure that the tools they deploy are accurate, safe and trusted, and that they help humans flourish.

Generative AI is evolving quickly, so the concrete steps businesses need to take will evolve over time. But sticking to a firm ethical framework can help organizations navigate this period of rapid transformation.

KEY TAKEAWAYS

- **Popularity and Potential of Generative AI:**

Generative AI has the potential to revolutionize various sectors, from customer interactions to business growth. A significant number of senior IT leaders are prioritizing its adoption. However, there are concerns about security risks and potential biased outcomes.

- **Ethical Implications:**

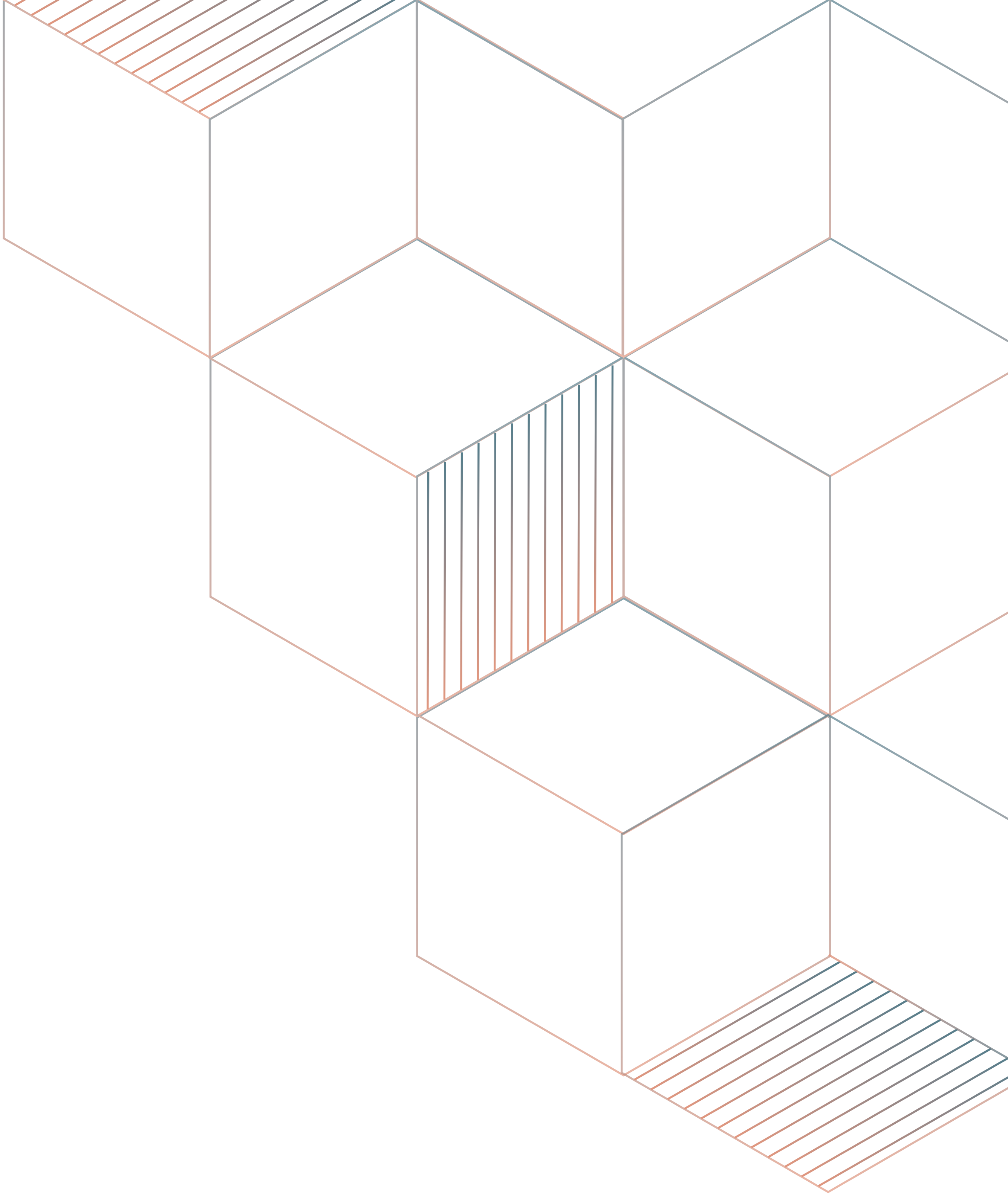
The use of generative AI in business settings differs from individual use. There are legal, financial, and ethical implications, especially if the generated content is inaccurate or offensive. The potential harm from incorrect AI-generated content can vary, emphasizing the need for clear ethical guidelines.

- **Guidelines for Ethical Development:**
  - **Accuracy:** Train models on reliable data and communicate uncertainties in responses.
  - **Safety:** Prioritize mitigating bias, ensuring data privacy, and conducting security assessments.
  - **Honesty:** Respect data provenance and be transparent when AI creates content.
  - **Empowerment:** AI should augment human capabilities, especially in sectors where trust is paramount.
  - **Sustainability:** Consider the environmental impact of training large AI models.

- **Integrating Generative AI:** Organizations should:
  - Use zero-party or first-party data for training.
  - Ensure data is up-to-date and well-labeled.
  - Involve humans in the review process to ensure accuracy and mitigate bias.
  - Continuously test AI models and seek feedback from various stakeholders.

- **Commitment to Ethical AI:**

As generative AI becomes mainstream, organizations have a responsibility to use it ethically. Adhering to a strong ethical framework can help businesses navigate the rapid transformations in the AI landscape.



REGULATION IS COMING: ETHICAL POLICYMAKING IN PRACTICE

How Do Foundation Models Comply with the EU AI Act? Grading LLMs

Stanford HAI | Contributors

Stanford University
Human-Centered
Artificial Intelligence

Percy Liang

Associate Professor of Computer Science

Daniel Zhang

Senior Manager for Policy Initiatives

Kevin Klyman

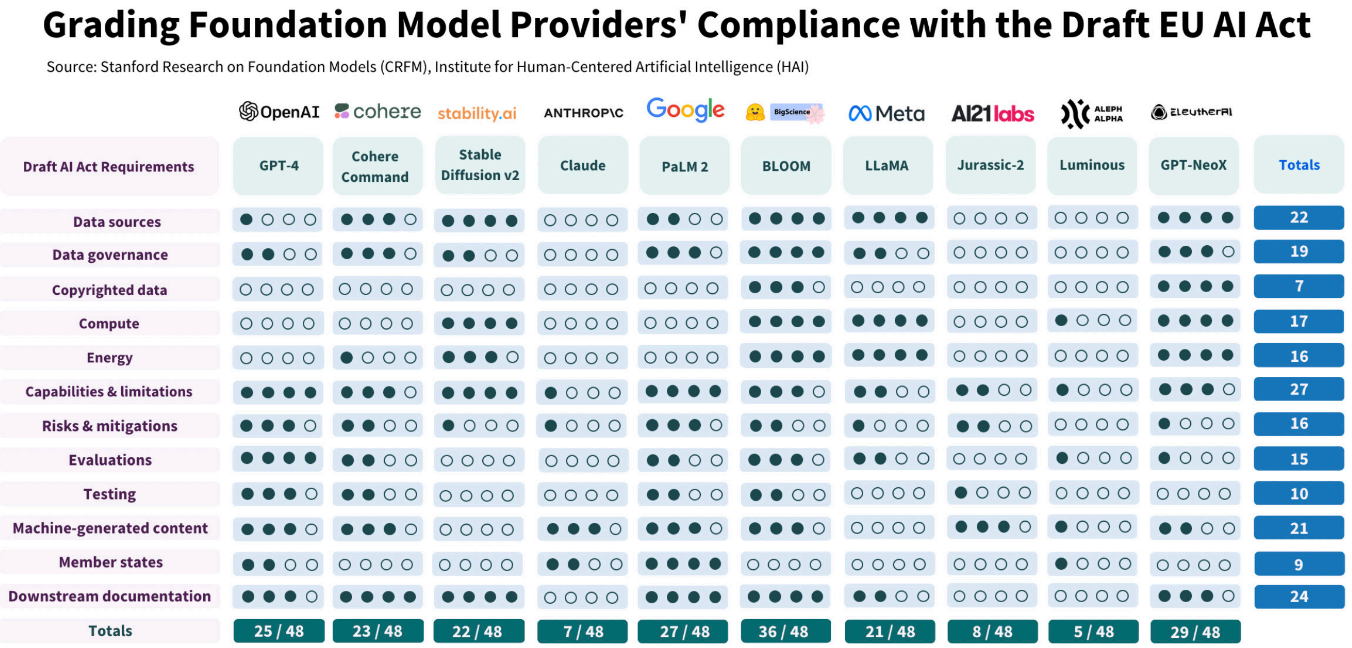
Technology Policy Strategist

Rishi Bommasani

Stanford PhD Candidate

How Do Foundation Models Comply with the EU AI Act? Grading LLMs

By Rishi Bommasani, Kevin Klyman, Daniel Zhang and Percy Liang



Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

	OpenAI	cohere	stability.ai	ANTHROPIC	Google	BigScience	Meta	AI21labs	ALEPH ALPHA	EleutherAI	
Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	Totals
Data sources	●○○○	●●●○	●●●●	○○○○	●●○○	●●●●	●●●●	○○○○	○○○○	●●●●	22
Data governance	●●○○	●●●○	●●○○	○○○○	●●●○	●●●●	●●○○	○○○○	○○○○	●●●○	19
Copyrighted data	○○○○	○○○○	○○○○	○○○○	○○○○	●●●○	○○○○	○○○○	○○○○	●●●●	7
Compute	○○○○	○○○○	●●●●	○○○○	○○○○	●●●●	●●●●	○○○○	●○○○	●●●●	17
Energy	○○○○	●○○○	●●●○	○○○○	○○○○	●●●●	●●●●	○○○○	○○○○	●●●●	16
Capabilities & limitations	●●●●	●●●○	●●●●	●○○○	●●●●	●●●○	●●○○	●●○○	●○○○	●●●○	27
Risks & mitigations	●●●○	●●○○	●○○○	●○○○	●●●○	●●○○	●○○○	●●○○	○○○○	●○○○	16
Evaluations	●●●●	●●○○	○○○○	○○○○	●●○○	●●●○	●●○○	○○○○	●○○○	●○○○	15
Testing	●●●○	●●○○	○○○○	○○○○	●●○○	●●○○	○○○○	●○○○	○○○○	○○○○	10
Machine-generated content	●●●○	●●●○	○○○○	●●●○	●●●○	●●●○	○○○○	●●●○	●○○○	●●○○	21
Member states	●●○○	○○○○	○○○○	●●○○	●●●●	○○○○	○○○○	○○○○	●○○○	○○○○	9
Downstream documentation	●●●○	●●●●	●●●●	○○○○	●●●●	●●●●	●●○○	○○○○	○○○○	●●●○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

Foundation models like ChatGPT are transforming society with their remarkable capabilities, serious risks, rapid deployment, unprecedented adoption, and unending controversy. Simultaneously, the European Union (EU) is finalizing its AI Act as the world's first comprehensive regulation to govern AI, and just yesterday the European Parliament adopted a draft of the Act by a vote of 499 in favor, 28 against, and 93 abstentions. The Act includes explicit obligations for foundation model providers like OpenAI and Google.

In this post, we evaluate whether major foundation model providers currently comply with these draft requirements and find that they largely do not. Foundation model providers rarely disclose adequate information regarding the data, compute, and deployment of their models as well as the key characteristics of the models themselves. In particular, foundation model providers generally do not comply with draft requirements to describe the use of copyrighted training data, the hardware used and emissions produced in training, and how they evaluate and test models. As a result, we recommend that policymakers prioritize transparency, informed by the AI Act's requirements. Our assessment demonstrates that it is currently feasible for foundation model providers to comply with the AI Act, and that disclosure related to foundation models' development, use, and performance would improve transparency in the entire ecosystem.

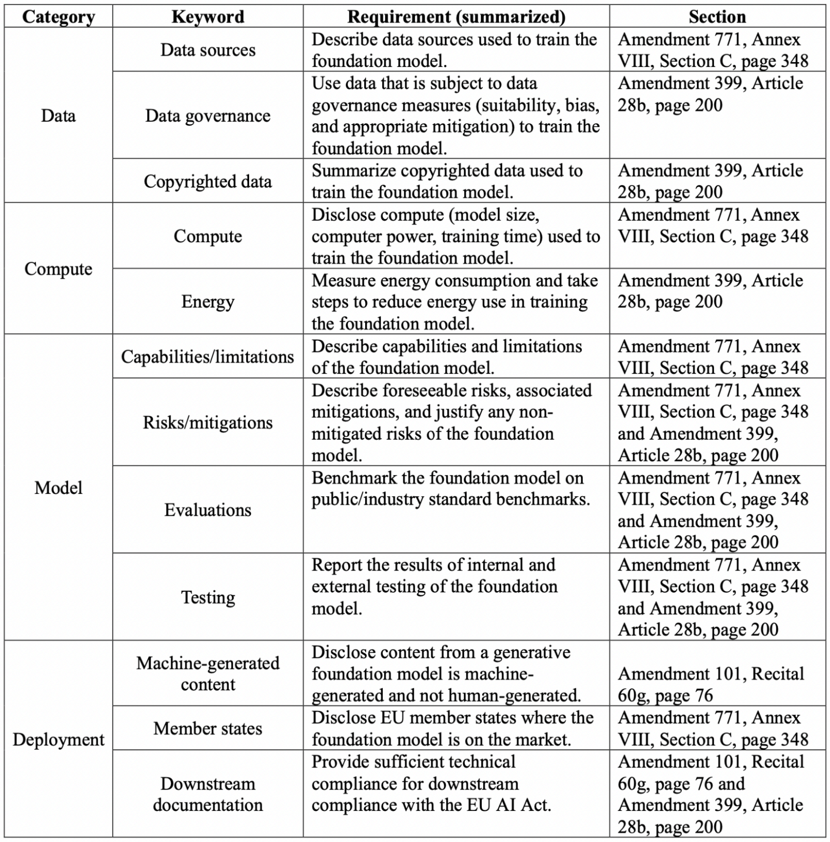
Motivation

Foundation models are at the center of global discourse on AI: the emerging technological paradigm has concrete and growing impact on the economy, policy, and society. In parallel, the EU AI Act is the most important regulatory initiative on AI in the world today. The Act will not only impose requirements for AI in the EU, a population of 450 million people, but also set precedent for AI regulation around the world (the Brussels effect). Policymakers across the globe are already drawing inspiration from the AI Act, and multinational companies may change their global practices to maintain a single AI development process. How we regulate foundation models will structure the broader digital supply chain and shape the technology's societal impact.

Our assessment establishes the facts about the status quo and motivates future intervention.

1. Status quo. What is the current conduct of foundation model providers? And, as a result, how will the EU AI Act (if enacted, obeyed, and enforced) change the status quo? We specifically focus on requirements where providers fall short at present.
2. Future intervention. For EU policymakers, where is the AI Act underspecified and where is it insufficient with respect to foundation models? For global policymakers, how should their priorities change based on our findings? And for foundation model providers, how should their business practices evolve to be more responsible? Overall, our research underscores that transparency should be the first priority to hold foundation model providers accountable.

Methodology



Category	Keyword	Requirement (summarized)	Section
Data	Data sources	Describe data sources used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Data governance	Use data that is subject to data governance measures (suitability, bias, and appropriate mitigation) to train the foundation model.	Amendment 399, Article 28b, page 200
	Copyrighted data	Summarize copyrighted data used to train the foundation model.	Amendment 399, Article 28b, page 200
Compute	Compute	Disclose compute (model size, computer power, training time) used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Energy	Measure energy consumption and take steps to reduce energy use in training the foundation model.	Amendment 399, Article 28b, page 200
Model	Capabilities/limitations	Describe capabilities and limitations of the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Risks/mitigations	Describe foreseeable risks, associated mitigations, and justify any non-mitigated risks of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
	Evaluations	Benchmark the foundation model on public/industry standard benchmarks.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
	Testing	Report the results of internal and external testing of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
Deployment	Machine-generated content	Disclose content from a generative foundation model is machine-generated and not human-generated.	Amendment 101, Recital 60g, page 76
	Member states	Disclose EU member states where the foundation model is on the market.	Amendment 771, Annex VIII, Section C, page 348
	Downstream documentation	Provide sufficient technical compliance for downstream compliance with the EU AI Act.	Amendment 101, Recital 60g, page 76 and Amendment 399, Article 28b, page 200

Table 1. We identify, categorize, summarize, and source requirements from the draft AI Act adopted by EU Parliament.

Below is a summary of our approach, including all relevant details in the referenced documents.

1. We extract 22 requirements directed towards foundation model providers from the European Parliament's version of the Act. We select 12 of the 22 requirements to assess—these requirements are able to be meaningfully evaluated using public information.

2. We categorize the 12 requirements as pertaining to (i) data resources (3), (ii) compute resources (2), (iii) the model itself (4), or (iv) deployment practices (3). Many of these requirements center on transparency: for example, disclosure of what data was used to train the foundation model, how the model performs on standard benchmarks, and where it is deployed. We summarize the 12 requirements in the table above.

3. We design a 5-point rubric for each of the 12 requirements. While the Act states high-level obligations should be interpreted or enforced. Our rubrics come from our expertise on the societal impact of

foundation models. These rubrics can directly inform statutory interpretation or standards, including in areas where the Act's language is especially unclear.

4. We assess the compliance of 10 foundation model providers—and their flagship foundation models—with 12 of the Act's requirements for foundation models based on our rubrics. The two lead authors independently scored all the providers for all requirements with substantial inter-annotator agreement of Cohen's Kappa = 0.74. We merge scores through panel discussion with all authors involved in this work. While comprehensive assessment of compliance with these requirements will require additional guidance from the EU, our research on providers' current practices will play a valuable role when regulators ultimately assess compliance.

Findings

We present the final scores in the above figure with the justification for every grade made available. Our results demonstrate a striking range in compliance across model providers: some providers score less than 25% (AI21 Labs, Aleph Alpha, Anthropic) and only one provider scores at least 75% (Hugging Face/BigScience) at present. Even for the highest-scoring providers, there is still significant margin for improvement. This confirms that the Act (if enacted, obeyed, and enforced) would yield significant change to the ecosystem, making substantial progress towards more transparency and accountability.

Persistent challenges. We see four areas where many organizations receive poor scores (generally 0 or 1 out of 4). They are (i) copyrighted data, (ii) compute/energy, (iii) risk mitigation, and (iv) evaluation/testing. These speak to established themes in the scientific literature:

- Unclear liability due to copyright. Few providers disclose any information about the copyright status of training data. Many foundation models are trained on data that is curated from the Internet, of which a sizable fraction is likely copyrighted. The legal validity of training on this data as a matter of fair use, especially for data with specific licenses, and of reproducing this, data remains unclear.
- Uneven reporting of energy use. Foundation model providers inconsistently report energy usage, emissions, their strategies for measurement of emissions, and any measures taken to mitigate emissions. How to measure the energy required to train foundation models is contentious (Strubell et al., 2019; Patterson et al., 2021). Regardless, the reporting of these costs proves to be unreliable, in spite of many efforts that have built tools to facilitate such reporting (Lacoste et al., 2019; Henderson et al., 2020; Luccioni et al., 2023).
- Inadequate disclosure of risk mitigation/non-mitigation. The risk landscape for foundation models is immense, spanning many forms of malicious use, unintentional harm, and structural or systemic risk(Bender et al., 2021; Bommasani et al., 2021; Weidinger et al., 2021). While many foundation model providers enumerate risks, relatively few disclose the mitigations they implement and the efficacy of these mitigations. The Act also requires that providers describe "non-mitigated risks with an explanation on the reason why they cannot be mitigated", which none of the providers we assess do.
- Absence of evaluation standards/auditing ecosystem. Foundation model providers rarely measure models' performance in terms of intentional harms such as malicious use or factors such as robustness and calibration. Many in the community have called for more evaluations, but standards for foundation model evaluation (especially beyond language models) remain a work-in-progress (Liang et al., 2022, Bommasani et al., 2023, Solaiman et al., 2023). In the U.S., the mandate for NIST to create AI testbeds under Section 10232 of the CHIPS and Science Act identifies one path towards such standards.

Open vs restricted/closed models. We find a clear dichotomy in compliance as a function of release strategy, or the extent to which foundation model providers make their models publicly available. At present, foundation model providers adopt a variety of release strategies, with no established norms. While release strategies are not binary and exist on a spectrum, for simplicity we consider broadly open releases (e.g. EleutherAI's GPT-NeoX, Hugging Face/BigScience's BLOOM, Meta's LLaMA) vs restricted/closed releases (e.g. Google's PaLM 2, OpenAI's GPT-4, Anthropic's Claude). Open releases generally achieve strong scores on resource disclosure requirements (both data and compute), with EleutherAI receiving 19/20 for these categories. However, such open releases make it challenging to monitor or control deployment, with more restricted/closed releases leading to better scores on deployment-related requirements. For instance, Google's PaLM 2 receives 11/12 for deployment. We emphasize that EU policymakers should consider strengthening deployment requirements for entities that bring foundation models to market to ensure there is sufficient accountability across the digital supply chain.

The relationship between release and area-specific compliance to some extent aligns with our intuitions. Open releases are often conducted by organizations that emphasize transparency, leading to a similar commitment to disclosing the resources required to build their foundation models. Restricted or closed releases, by contrast, often coincide with models that power the provider's flagship products and services, meaning the resources underlying the model may be seen as a competitive advantage (e.g. data decomposition) or a liability (e.g. copyrighted data). In addition, open-sourcing a model makes it much more difficult to monitor or influence downstream use, whereas APIs or developer-mediated access provide easier means for structured access.

Overall feasibility of compliance. No foundation model provider achieves a perfect score, with ample room for improvement in most cases. Therefore, we consider whether it is currently feasible for organizations to fully comply with all requirements. While we believe that with sufficient incentives (e.g. fines for noncompliance) companies will change their conduct, even in the absence of strong regulatory pressure, many providers could reach total scores in the high 30s or 40s through meaningful, but plausible, changes. To be concrete, the entry-wise maximum across OpenAI and Hugging Face/ BigScience is 42 (almost 90% compliance). We conclude that enforcing these 12 requirements in the Act would bring substantive change while remaining within reach for providers.

Releases of foundation models have generally become less transparent, as evidenced by major releases in recent months. The reports for OpenAI's GPT-4 and Google's PaLM 2 openly state that they do not report many relevant aspects about data and compute. The GPT-4 paper reads "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."

We believe sufficient transparency to satisfy the Act's requirements related to data, compute and other factors should be commercially feasible if foundation model providers collectively take action as the result of industry standards or regulation. We see no significant barriers that would prevent every provider from improving how it discusses limitations and risks as well as reporting on standard benchmarks. Although open-sourcing may make aspects of deployment disclosure challenging, feasible improvements in disclosure of machine-generated content or availability of downstream documentation abound. While progress in each of these areas requires some work, in many cases we believe this work is minimal relative to building and providing the foundation model and should be seen as a prerequisite for being a responsible and reputable model provider.

Recommendations

We direct our recommendations to three parties: (i) EU policymakers working on the EU AI Act, (ii) global policymakers working on AI policy, and (iii) foundation model providers operating across the ecosystem.

EU policymakers.

- The implementation of the EU AI Act and technical standards to follow should specify areas of the Act that are underspecified. Given our expertise in evaluations, we emphasize the importance of specifying which dimensions of performance are necessary to disclose to comply with the mandate for a "description of the model's performance". We advocate that several factors such as accuracy, robustness, fairness, and efficiency be considered necessary for compliance (the NIST AI Risk Management Framework provides a similar list).
- The EU AI Act should consider additional critical factors to ensure adequate transparency and accountability of foundation model providers, including the disclosure of usage patterns: such requirements would mirror transparency reporting for online platforms, the lack of which has been a chronic inhibitor for effective platform policy. We need to understand how foundation models are used (e.g. for providing medical advice, preparing legal documents) to hold their providers to account. We encourage policymakers to consider making these requirements apply only to the most influential foundation model providers, directly mirroring how the EU's Digital Services Act places special requirements on Very Large Online Platforms to avoid overburdening smaller companies.
- For effective enforcement of the EU AI Act to change the conduct of the powerful organizations that build foundation models, the EU must make requisite technical resources and talent available to enforcement agencies, especially given the broader AI auditing ecosystem envisioned in the Act. Our assessment process made clear that technical expertise on foundation models is necessary to understand this complex ecosystem.

Global policymakers.

- Transparency should be the first priority for policy efforts: it is an essential precondition for rigorous science, sustained innovation, accountable technology, and effective regulation. Our work shows transparency is uneven at present, and an area where the EU AI Act will bring clear change that policy elsewhere should match. The history of social media regulation provides clear lessons for policymakers—failing to ensure sufficient platform transparency led to many of the harms of social media; we should not reproduce these failures for the next transformational technology in foundation models.
- Disclosure of copyrighted training data is the area where we find foundation model providers achieve the worst compliance. Legislators, regulators and courts should clarify how copyright relates to (i) the training procedure, including the conditions under which copyright or licenses must be respected during training as well as the measures model providers should take to reduce the risk of copyright infringement and (ii) the output of generative models, including the conditions under which machine-generated content infringes on the rights of content creators in the same market.

Foundation model providers.

- Our work indicates where each foundation model provider can improve. We highlight many steps that are low-hanging fruit, such as improving the documentation made available to downstream developers that build on foundation models. In various cases, some providers score worse than others with similar release strategies (e.g. different providers that deploy their foundation models via an API). Therefore, providers can and should improve compliance by emulating similar providers that are best-in-class.

- Foundation model providers should work towards industry standards that will help the overall ecosystem become more transparent and accountable. The standards-setting process should involve stakeholders beyond foundation model providers with specific attention towards parties that can better represent the public interest like academia and civil society.

Limitations

While we offer expertise on foundation models, our reading of the draft law is not genuine statutory interpretation, though it could inform such interpretation (especially where the law is unclear). The AI Act remains under discussion and will be finalized during the upcoming trilogue between the EU Commission, Council, and Parliament. Foundation model providers also have requirements under provisions of the AI Act that do not address only foundation models, such as when their foundation models are integrated into high-risk AI systems. Therefore, our assessments might diverge from foundation models providers' compliance with the final version of the AI Act. Given that our assessment is based on, and limited by, publicly available information, we encourage foundation model providers to provide feedback to us and respond to these scores.

Conclusion

We find that foundation model providers unevenly comply with the stated requirements of the draft EU AI Act. Enacting and enforcing the EU AI Act will bring about significant positive change in the foundation model ecosystem. Foundation model providers' compliance with requirements regarding copyright, energy, risk, and evaluation is especially poor, indicating areas where model providers can improve. Our assessment shows sharp divides along the boundary of open vs. closed releases: we believe that all providers can feasibly improve their conduct, independent of where they fall along this spectrum. Overall, our analysis speaks to a broader trend of waning transparency: providers should take action to collectively set industry standards that improve transparency, and policymakers should take action to ensure adequate transparency underlies this general-purpose technology. This work is just the start of a broader initiative at the Center for Research on Foundation Models to directly assess and improve the transparency of foundation model providers, complementing our efforts on holistic evaluation, ecosystem documentation, norms development, policy briefs, and policy recommendations.

Acknowledgments

We thank Alex Engler, Arvind Narayanan, Ashwin Ramaswami, Dan Ho, Irene Solaiman, Marietje Schaake, Peter Cihon, and Sayash Kapoor for feedback on this effort. We thank Alex Engler, Arvind Narayanan, Ashwin Ramaswami, Aviv Ovadya, Conor Griffin, Dan Ho, Iason Gabriel, Irene Solaiman, Joslyn Barnhart, Markus Anderljung, Sayash Kapoor, Seliem El-Sayed, Seth Lazar, Stella Bidermann, Toby Shevlane, and Zak Rogoff for broader discussions on this topic. We thank Madeleine Wright for designing the graphics.

KEY TAKEAWAYS

- **Foundation Models and the EU AI Act:** The European Union (EU) is in the process of finalizing its AI Act, which is set to be the world's first comprehensive regulation governing AI. This Act has explicit obligations for foundation model providers like OpenAI and Google. However, many of these providers currently do not comply with the draft requirements, especially in areas like data disclosure, compute resources, and model deployment.

- **Major Gaps in Compliance:** The article reveals that foundation model providers often do not disclose adequate information about their models' data, compute resources, deployment, and other key characteristics. Specific areas of non-compliance include the use of copyrighted training data, the hardware and emissions involved in training, and the evaluation and testing of models.

- **Persistent Challenges:**

Copyrighted Data: Few providers disclose information about the copyright status of their training data, raising concerns about the legality of using such data.

Energy Reporting: There's inconsistent reporting on energy usage and emissions during model training. Risk Mitigation: While many providers identify potential risks, few disclose the measures they've taken to mitigate them.

Evaluation Standards: There's a lack of standardized evaluation measures for foundation models, especially in areas like robustness and calibration.

Open vs. Restricted Models: The article identifies a clear distinction in compliance based on the release strategy of the models. Open releases, like those from EleutherAI, tend to score higher on resource disclosure but face challenges in monitoring deployment. In contrast, restricted releases, like Google's PaLM 2, score better on deployment-related requirements.

Feasibility of Compliance: The article suggests that while no foundation model provider currently achieves full compliance, it is feasible for them to do so. With the right incentives, such as potential fines for non-compliance, providers could make significant improvements.

- **Recommendations:**

For EU Policymakers: The article suggests that the EU AI Act should be more specific in areas that are currently underspecified and should consider additional factors to ensure transparency and accountability.

For Global Policymakers: Emphasizing transparency should be a top priority, and there should be clarity on how copyright relates to AI training and outputs.

For Foundation Model Providers: Providers should aim to improve their practices, especially in areas where they currently fall short. The article also recommends the development of industry standards to enhance transparency and accountability.

European Parliament Research Service: The EU’s AI Act

Tambiama Madiega

Biography

Tambiama Madiega is a seasoned Policy Analyst at the European Parliament, where he’s been contributing for over 8 years, focusing on digital policies including intellectual property rights, platform regulation, e-commerce, telecom, and artificial intelligence. Before this, he played pivotal roles at the European Commission, leading regulatory coordination and market analysis in the telecoms and ICT sectors. Tambiama’s expertise also extends to the private sector, having worked with SITA on telecoms and air transport regulations and as an Associate Lawyer specializing in ICT laws at Hogan & Hartson. He’s an alumnus of the European University Institute with a PhD in Competition Law and Telecoms Regulation. Tambiama also holds credentials from the University of Columbia, University of California at Berkeley, McGill University, and Université Paris Nanterre.

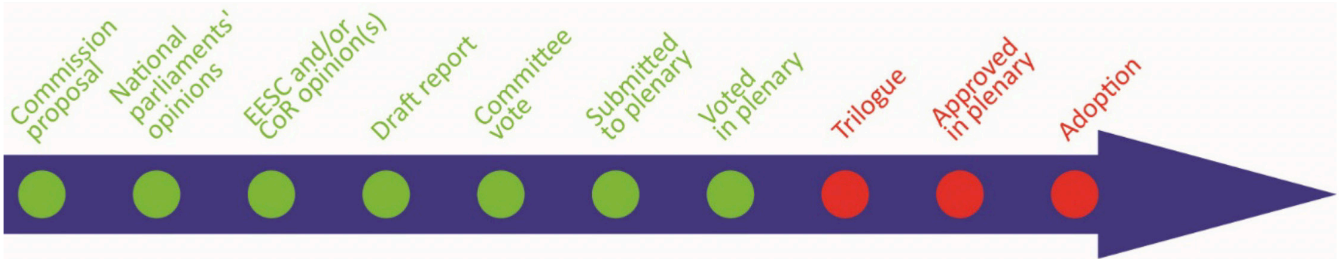
European Parliament Research Service: The EU's AI Act

By Tambiama Madiega

Overview

The European Commission tabled a proposal for an EU regulatory framework on artificial intelligence (AI) in April 2021. The draft AI act is the first ever attempt to enact a horizontal regulation for AI. The proposed legal framework focuses on the specific utilisation of AI systems and associated risks. The Commission proposes to establish a technology-neutral definition of AI systems in EU law and to lay down a classification for AI systems with different requirements and obligations tailored on a 'risk-based approach'. Some AI systems presenting 'unacceptable' risks would be prohibited. A wide range of 'high-risk' AI systems would be authorised, but subject to a set of requirements and obligations to gain access to the EU market. Those AI systems presenting only 'limited risk' would be subject to very light transparency obligations. The Council agreed the EU Member States' general position in December 2021. Parliament voted on its position in June 2023. EU lawmakers are now starting negotiations to finalise the new legislation, with substantial amendments to the Commission's proposal including revising the definition of AI systems, broadening the list of prohibited AI systems, and imposing obligations on general purpose AI and generative AI models such as ChatGPT.

Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts		
Committees responsible:	Internal Market and Consumer Protection (IMCO) and Civil Liberties, Justice and Home Affairs (LIBE) (jointly under Rule 58)	COM(2021)206 21.4.2021 2021/0106(COD)
Rapporteurs:	Brando Benifei (S&D, Italy) and Dragoş Tudorache (Renew, Romania)	
Shadow rapporteurs:	Deirdre Clune, Axel Voss (EPP); Petar Vitanov (S&D); Svenja Hahn, (Renew); Sergey Lagodinsky, Kim Van Sparrentak (Greens/EFA); Rob Rooken, Kosma Złotowski (ECR); Jean-Lin Lacapelle, Jaak Madison (ID); Cornelia Ernst, Kateřina Konecna (The Left)	Ordinary legislative procedure (COD) (Parliament and Council on equal footing – formerly 'co-decision')
Next steps expected:	Trilogue negotiations	



Introduction

AI technologies are expected to bring a wide array of economic and societal benefits to a wide range of sectors, including environment and health, the public sector, finance, mobility, home affairs and agriculture. They are particularly useful for improving prediction, for optimising operations and resource allocation, and for personalising services. However, the implications of AI systems for fundamental rights protected under the EU Charter of Fundamental Rights, as well as the safety risks for users when AI technologies are embedded in products and services, are raising concern. Most notably, AI systems may jeopardise fundamental rights such as the right to non-discrimination, freedom of expression, human dignity, personal data protection and privacy. Given the fast development of these technologies, in recent years AI regulation has become a central policy question in the European Union (EU). Policy-makers pledged to develop a 'humancentric' approach to AI to ensure that Europeans can benefit from new technologies developed and functioning according to the EU's values and principles. In its 2020 White Paper on Artificial Intelligence, the European Commission committed to promote the uptake of AI and address the risks associated with certain uses of this new technology. While the European Commission initially adopted a soft-law approach, with the publication of its non-binding 2019 Ethics Guidelines for Trustworthy AI and Policy and investment recommendations, it has since shifted towards a legislative approach, calling for the adoption of harmonised rules for the development, placing on the market and use of AI systems.

AI regulatory approach in the world. While the United States of America (USA) had initially taken a lenient approach towards AI, calls for regulation have recently been mounting. The Cyberspace Administration of China is also consulting on a proposal to regulate AI, while the UK is working on a set of pro-innovation regulatory principles. At international level, the Organisation for Economic Co-operation and Development (OECD) adopted a (non-binding) Recommendation on AI in 2019, UNESCO adopted Recommendations on the Ethics of AI in 2021, and the Council of Europe is currently working on an international convention on AI. Furthermore, in the context of the newly established EU-US tech partnership (the Trade and Technology Council), the EU and USA are seeking to develop a mutual understanding on the principles underlying trustworthy and responsible AI. EU lawmakers issued a joint statement in May 2023 urging President Biden and European Commission President Ursula von der Leyen to convene a summit to find ways to control the development of advanced AI systems such as ChatGPT.

Parliament's Starting Position

Leading the EU-level debate, the European Parliament called on the European Commission to assess the impact of AI and to draft an EU framework for AI, in its wide-ranging 2017 recommendations on civil law rules on robotics. More recently, in 2020 and 2021, the Parliament adopted a number of non-legislative resolutions calling for EU action, as well as two legislative resolutions calling for the adoption of EU legislation in the field of AI. A first legislative resolution asked that the Commission establish a legal framework of ethical principles for the development, deployment and use of AI, robotics and related technologies in the Union. A second legislative resolution called for harmonisation of the legal framework for civil liability claims and imposition of a regime of strict liability on operators of high-risk AI systems. Furthermore, the Parliament adopted a series of recommendations calling for a common EU approach to AI in the intellectual property, criminal law, education, culture and audiovisual areas, and regarding civil and military AI uses.

Council Starting Position

In the past, the Council has repeatedly called for the adoption of common AI rules, including in 2017 and 2019. More recently, in 2020, the Council called upon the Commission to put forward concrete proposals that take existing legislation into account and follow a risk-based, proportionate and, if necessary, regulatory approach. Furthermore, the Council invited the EU and the Member States to consider effective measures for identifying, predicting and responding to the potential impacts of digital technologies, including AI, on fundamental rights.

Preparation of the Proposal

Following the White Paper on Artificial Intelligence adopted in February 2020, the Commission launched a broad public consultation in 2020 and published an Impact Assessment of the regulation on artificial intelligence, a supporting study and a draft proposal, which received feedback from a variety of stakeholders. In its impact assessment, the Commission identifies several problems raised by the development and use of AI systems, due to their specific characteristics.

The Changes the Proposal Would Bring

The draft AI act has been designed as a horizontal EU legislative instrument applicable to all AI systems placed on the market or used in the Union.

Purpose, legal basis and scope

The general objective of the proposed AI act unveiled in April 2021 is to ensure the proper functioning of the single market by creating the conditions for the development and use of trustworthy AI systems in the Union. The draft sets out a harmonised legal framework for the development, placing on the Union market, and the use of AI products and services. In addition, the AI act proposal seeks to achieve a set of specific objectives: (i) ensure that AI systems placed on the EU market are safe and respect existing EU law, (ii) ensure legal certainty to facilitate investment and innovation in AI, (iii) enhance governance and effective enforcement of EU law on fundamental rights and safety requirements applicable to AI systems, and (iv) facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.

The new AI framework, based on Article 114 and Article 16 of the Treaty on the Functioning of the European Union (TFEU), would enshrine a technology-neutral definition of AI systems and adopt a risk-based approach, which lays down different requirements and obligations for the development, placing on the market and use of AI systems in the EU. In practice, the proposal defines common mandatory requirements applicable to the design and development of AI systems before they are placed on the market and harmonises the way ex-post controls are conducted. The proposed AI act would complement existing and forthcoming, horizontal and sectoral EU safety regulation. The Commission proposes to follow the logic of the new legislative framework (NLF), i.e. the EU approach to ensuring a range of products comply with the applicable legislation when they are placed on the EU market through conformity assessments and the use of CE marking.

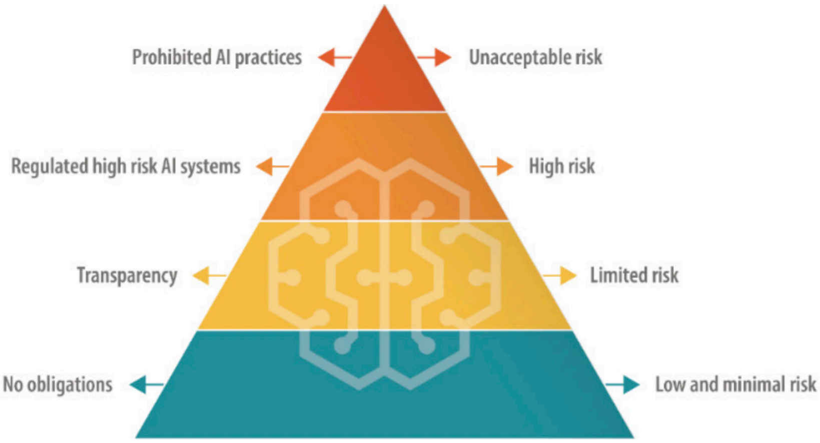
The new rules would apply primarily to providers of AI systems established within the EU or in a third country placing AI systems on the EU market or putting them into service in the EU, as well as to users of AI systems located in the EU. To prevent circumvention of the regulation, the new rules would also apply to providers and users of AI systems located in a third country where the output produced by those systems is used in the EU. However, the draft regulation does not apply to AI systems developed or used exclusively for military purposes, to public authorities in a third country, nor to international organisations, or authorities using AI systems in the framework of international agreements for law enforcement and judicial cooperation.

Definitions

No single definition of artificial intelligence is accepted by the scientific community and the term 'AI' is often used as a 'blanket term' for various computer applications based on different techniques, which exhibit capabilities commonly and currently associated with human intelligence. The High Level Expert Group on AI proposed a baseline definition of AI that is increasingly used in the scientific literature, and the Joint Research Centre has established an operational definition of AI based on a taxonomy that maps all the AI subdomains from a political, research and industrial perspective. However, the Commission found that the notion of an AI system should be more clearly defined, given that the determination of what an 'AI system' constitutes is crucial for the allocation of legal responsibilities under the new AI framework. The Commission therefore proposes to establish a legal definition of 'AI system' in EU law, which is largely based on a definition already used by the OECD. Article 3(1) of the draft act states that 'artificial intelligence system' means

...software that is developed with [specific] techniques and approaches [listed in Annex 1] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.

Annex 1 of the proposal lays out a **list of techniques and approaches** that are used today to develop AI. Accordingly, the notion of 'AI system' would refer to a range of software-based technologies that encompasses **'machine learning'**, **'logic and knowledge-based'** systems, and **'statistical'** approaches. This broad definition covers AI systems that can be used on a stand-alone basis or as a component of a product. Furthermore, the proposed legislation aims to be future-proof and cover current and future AI technological developments. To that end, the Commission would complement the Annex 1 list with new approaches and techniques used to develop AI systems as they emerge – through the adoption of delegated acts (Article 4). Furthermore, Article 3 provides a long list of definitions including that of 'provider' and 'user' of AI systems (covering both public and private entities), as well as 'importer' and 'distributor', 'emotion recognition', and 'biometric categorisation'.



Source: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Furthermore, Article 3 provides a long list of definitions including that of 'provider' and 'user' of AI systems (covering both public and private entities), as well as 'importer' and 'distributor', 'emotion recognition', and 'biometric categorisation'.

Risk-based approach

The use of AI, with its specific characteristics (e.g. opacity, complexity, dependency on data, autonomous behaviour), can adversely affect a number of fundamental rights and users' safety. To address those concerns, the draft AI act follows a risk-based approach whereby legal intervention is tailored to a concrete level of risk. To that end, the draft AI act distinguishes between AI systems posing (i) unacceptable risk, (ii) high risk, (iii) limited risk, and (iv) low or minimal risk. AI applications would be regulated only as strictly necessary to address specific levels of risk.

1. Unacceptable risk: Prohibited AI practices

Title II (Article 5) of the proposed AI act explicitly bans harmful AI practices that are considered to be a clear threat to people's safety, livelihoods and rights, because of the 'unacceptable risk' they create. Accordingly, it would be prohibited to place on the market, put into services or use in the EU:

- AI systems that deploy harmful manipulative 'subliminal techniques';
- AI systems that exploit specific vulnerable groups (physical or mental disability);
- AI systems used by public authorities, or on their behalf, for social scoring purposes;
- 'Real-time' remote biometric identification systems in publicly accessible spaces for law enforcement purposes, except in a limited number of cases.

2. High risk: Regulated high-risk AI systems

Title III (Article 6) of the proposed AI act regulates 'high-risk' AI systems that create adverse impact on people's safety or their fundamental rights. The draft text distinguishes between two categories of high-risk AI systems.

- Systems used as a safety component of a product or falling under EU health and safety harmonisation legislation (e.g. toys, aviation, cars, medical devices, lifts).
- Systems deployed in eight specific areas identified in Annex III, which the Commission could update as necessary through delegated acts (Article 7):
    - Biometric identification and categorisation of natural persons; o Management and operation of critical infrastructure
    - Education and vocational training
    - Employment, worker management and access to self-employment
    - Access to and enjoyment of essential private services and public services and benefits
    - Law enforcement
    - Migration, asylum and border control management
    - Administration of justice and democratic processes

All of these high-risk AI systems would be subject to a set of new rules including:
Requirement for an ex-ante conformity assessment: Providers of high-risk AI systems would be required to register their systems in an EU-wide database managed by the Commission before placing them on the market or putting them into service. Any AI products and services governed by existing product safety legislation will fall under the existing third-party conformity frameworks that already apply (e.g. for medical devices). Providers of AI systems not currently governed by EU legislation would have to conduct their own conformity assessment (self-assessment) showing that they comply with the new requirements and can use CE marking. Only high-risk AI systems used for biometric identification would require a conformity assessment by a 'notified body'.

Other requirements: Such high-risk AI systems would have to comply with a range of requirements particularly on risk management, testing, technical robustness, data training and data governance, transparency, human oversight, and cybersecurity (Articles 8 to 15). In this regard, providers, importers, distributors and users of high-risk AI systems would have to fulfil a range of obligations. Providers from outside the EU will require an authorised representative in the EU to (inter alia), ensure the conformity assessment, establish a post-market monitoring system and take corrective action as needed. AI systems that conform to the new harmonised EU standards, currently under development, would benefit from a presumption of conformity with the draft AI act requirements.

3. Limited risk: Transparency obligations

AI systems presenting 'limited risk', such as systems that interacts with humans (i.e. chatbots), emotion recognition systems, biometric categorisation systems, and AI systems that generate or manipulate image, audio or video content (i.e. deepfakes), would be subject to a limited set of transparency obligations.

4. Low or minimal risk: No obligations
All other AI systems presenting only low or minimal risk could be developed and used in the EU without conforming to any additional legal obligations. However, the proposed AI act envisages the creation of codes of conduct to encourage providers of non-high-risk AI systems to voluntarily apply the mandatory requirements for high-risk AI systems.

Governance, enforcement and sanctions
The proposal requires Member States to designate one or more competent authorities, including a national supervisory authority, which would be tasked with supervising the application and implementation of the regulation, and establishes a European Artificial Intelligence Board (composed of representatives from the Member States and the Commission) at EU level. National market surveillance authorities would be responsible for assessing operators' compliance with the obligations and requirements for high-risk AI systems. They would have access to confidential information (including the source code of the AI systems) and subject to binding confidentiality obligations. Furthermore, they would be required to take any corrective measures to prohibit, restrict, withdraw or recall AI systems that do not comply with the AI act, or that, although compliant, present a risk to health or safety of persons or to fundamental rights or other public interest protection. In case of persistent non-compliance, Member States will have to take all appropriate measures to restrict, prohibit, recall or withdraw the high-risk AI system at stake from the market. Administrative fines of varying scales (up to €30 million or 6 % of the total worldwide annual turnover), depending on the severity of the infringement, are set as sanctions for non-compliance with the AI act. Member States would need to lay down rules on penalties, including administrative fines and take all measures necessary to ensure that they are properly and effectively enforced.

Measures to support innovation
The Commission proposes that Member States, or the European Data Protection Supervisor, could establish a regulatory sandbox, i.e. a controlled environment that facilitates the development, testing and validation of innovative AI systems (for a limited period of time) before they are put on the market. Sandboxing will enable participants to use personal data to foster AI innovation, without prejudice to the GDPR requirements. Other measures are tailored specifically to small-scale providers and start-ups.

Advisory Committees

The European Economic and Social Committee adopted its opinion on the proposed artificial intelligence act on 22 September 2021.

National Parliaments

The deadline for the submission of reasoned opinions on the grounds of subsidiarity was 2 September 2021. Contributions were received from the Czech Chamber of Deputies and the Czech Senate, the Portuguese Parliament, the Polish Senate and the German Bundesrat.

Stakeholder Views

Definitions
Definitions are a contentious point of discussion among stakeholders. The Big Data Value Association, an industry-driven international not–for-profit organisation, stresses that the definition of AI systems

is quite broad and would cover far more than what is subjectively understood as AI, including the simplest search, sorting and routing algorithms, which would consequently be subject to new rules. Furthermore, they ask for clarification of how components of larger AI systems (such as pre-trained AI components from other manufacturers or components not released separately), should be treated. AmCham, the American Chamber of Commerce in the EU, suggests avoiding over-regulation by adopting a narrower definition of AI systems, focusing strictly on high-risk AI applications (and not extended to AI applications that are not high-risk, or software in general). AccessNow, an association defending users' digital rights, argues the definitions of 'emotion recognition' and 'biometric categorisation' are technically flawed, and recommends adjustments.

Risk-based approach

While they generally welcome the proposed AI act's risk-based approach, some stakeholders support wider prohibition and regulation of AI systems. Civil rights organisations call for a ban on indiscriminate or arbitrarily targeted use of biometrics in public or publicly accessible spaces, and for restrictions on the uses of AI systems, including for border control and predictive policing. AccessNow argues that the provisions concerning prohibited AI practices (Article 5) are too vague, and proposes a wider ban on the use of AI to categorise people based on physiological, behavioural or biometric data, for emotion recognition, as well as dangerous uses in the context of policing, migration, asylum, and border management. Furthermore, they call forstronger impact assessment and transparency requirements.

The European Enterprises Alliance stresses that there is general uncertainty about the roles and responsibilities of the different actors in the AI value chain (developers, providers, and users of AI systems). This is particularly challenging for companies providing general purpose application programming interfaces or open-source AI models that are not specifically intended for high-risk AI systems but are nevertheless used by third parties in a manner that could be considered high-risk. They also call for 'high-risk' to be redefined, based on the measurable harm and potential impact. AlgorithmWatch underlines that the applicability of specific rules should not depend on the type of technology, but on the impact it has on individuals and society. They call for the new rules to be defined according to the impact of the AI systems and recommend that every operator should conduct an impact assessment that assesses the system's risk levels on a case-by-case basis. Climate Change AI calls for climate change mitigation and adaptation to be taken into account in the classification rules for high-risk AI systems and impose environmental protection requirements.

Consumer protection

The European Consumer Organisation, BEUC, stresses that the proposal requires substantial improvement to guarantee consumer protection. The organisation argues that the proposal should have a broader scope and impose basic principles and obligations (e.g. on fairness, accountability and transparency) upon all AI systems, as well as prohibiting more comprehensively harmful practices (such as private entities' use of social scoring and of remote biometric identification systems in public spaces). Furthermore, consumers should be granted a strong set of rights, effective remedies and redress mechanisms, including collective redress.

Impact on investments and SMEs

There are opposing views on the impact of the proposed regulation on investment. A study by the Centre for Data Innovation (representing large online platforms) highlights that the compliance costs incurred under the proposed AI act would likely provoke a chilling effect on investment in AI in Europe, and could particularly deter small and medium-sized enterprises (SMEs) from developing high-risk AI systems. According to the Centre for Data Innovation, the AI act would cost the European economy €31 billion over the next five years and reduce AI investments by almost 20 %. However, such estimates of the compliance costs are challenged by the experts from the Centre for European

Policy Studies, as well as by other economists. The European Digital SME Alliance warns against overly stringent conformity requirements, asks for effective representation of SMEs in the standards-setting procedures and for making sandboxes mandatory in all EU Member States.

Academic and Other Views

While generally supporting the Commission's proposal, critics call for amendments, including revising the 'AI systems' definition, ensuring a better allocation of responsibility, strengthening enforcement mechanisms and fostering democratic participation. Among the main issues are:

AI systems definition

The legal definition of 'AI systems' contained in the proposed AI act has been heavily criticised. Smuha and others warn the definition lacks clarity and may lead to legal uncertainty, especially for some systems that would not qualify as AI systems under the draft text, while their use may have an adverse impact on fundamental rights. To address this issue, the authors propose to broaden the scope of the legislation to explicitly include all computational systems used in the identified highrisk domains, regardless of whether they are considered to be AI. According to the authors, the advantage would be in making application of the new rules more dependent on the domain in which the technology is used and the fundamental rights-related risks, rather than on a specific computational technique. Ebers and others consider that the scope of 'AI systems' is overly broad, which may lead to legal uncertainty for developers, operators, and users of AI systems and ultimately to over-regulation. They call on EU law-makers to exempt AI systems developed and used for research purposes and open-source software (OSS) from regulation. Other commentators question whether the proposed definition of 'AI systems' is truly technology neutral as it refers primarily to 'software', omitting potential future AI developments.

Risk-based approach

Academics also call for amendments, warning that the risk-based approach proposed by the Commission would not ensure a high level of protection of fundamental rights. Smuha and others argue that the proposal does not always accurately recognise the wrongs and harms associated with different kinds of AI systems and therefore does not appropriately allocate responsibility. Among other things, they recommend adding a procedure that enables the Commission to broaden the list of prohibited AI systems, and propose banning existing manipulative AI systems (e.g. deepfakes), social scoring and some biometrics. Ebers and others call for a more detailed classification of risks to facilitate industry self-assessment and support, as well as prohibiting more AI systems (e.g. biometrics), including in the context of private use. Furthermore, some highlight that the draft legislation does not address systemic sustainability risks created by AI especially in the area of climate and environmental protection.

Experts seem particularly concerned by the implementation of Article 5 (prohibited practices) and Article 6 (regulated high-risk practices). One of the major concerns raised is that the rules on prohibited and high-risk practices may prove ineffective in practice, because the risk assessment is left to provider self-assessment. Veale and Zuiderveen Borgesius warn that most providers can arbitrarily classify most high-risk systems as adhering to the rules using self-assessment procedures alone. Smuha and others recommend exploring whether certain high-risk systems would not benefit from a conformity assessment carried out by an independent entity prior to their deployment.
democratic oversight of the standardisation process.

Governance structure and enforcement and redress mechanisms

Ebers and others stress that the AI act lacks effective enforcement structures, as the Commission proposes to leave the preliminary risk assessment, including the qualification as high-risk, to the providers' self-assessment. They also raise concerns about the excessive delegation of regulatory power to private European standardisation organisations (ESOs), due to the lack of democratic oversight, the impossibility for stakeholders (civil society organisations, consumer associations) to influence the development of standards, and the lack of judicial means to control them once they have been adopted. Instead, they recommend that the AI act codifies a set of legally binding requirements for high-risk AI systems (e.g. prohibited forms of algorithmic discrimination), which ESOs may specify through harmonised standards. Furthermore, they advocate that European policymakers should strengthen democratic oversight of the standardisation process.

Commentators deplore a crucial gap in the AI act, which does not provide for individual enforcement rights. Ebers and others stress that individuals affected by AI systems and civil rights organisations have no right to complain to market surveillance authorities or to sue a provider or user for failure to comply with the requirements. Similarly, Veale and Zuiderveen Borgesius warn that, while some provisions of the draft legislation aim to impose obligations on AI systems users, there is no mechanism for complaint or judicial redress available to them. Smuha and others recommend amending the proposal to include, inter alia, an explicit right of redress for individuals and rights of consultation and participation for EU citizens regarding the decision to amend the list of high-risk systems in Annex III.

It has also been stressed that the text as it stands lacks proper coordination mechanisms between authorities, in particular concerning cross-border infringement. Consequently, the competence of the relevant authorities at national level should be clarified. Furthermore, guidance would be desirable on how to ensure compliance with transparency and information requirements, while simultaneously protecting intellectual property rights and trade secrets (e.g. to what extent the source code must be disclosed), not least to avoid diverging practices in the Member States.

Legislative process

The Council adopted its common position in December 2022. The Council's proposes, inter alia to:

- Narrow the definition of AI systems to systems developed through machine learning approaches and logic- and knowledge-based approaches
- Extend to private actors the prohibition on using AI for social scoring, and add cases when the use of 'real-time' remote biometric identification systems in publicly accessible spaces could exceptionally be allowed
- Impose requirements on general purpose AI systems by means of implementing acts
- Add new provisions to take into account situations where AI systems can be used for many different purposes (general purpose AI)
- Simplify the compliance framework for the AI Act and strengthen, in particular, the role of the AI Board

In Parliament, the file was assigned jointly (under Rule 58) to the Committee on Internal Market and Consumer Protection (IMCO) and the Committee on Civil Liberties, Justice and Home Affairs (LIBE), with Brando Benifei (S&D, Italy) and Dragos Tudorache, Renew, Romania) appointed as rapporteurs. In addition, the Legal Affairs Committee (JURI), the Committee on Industry, Research and Energy (ITRE) and the Committee on Culture and Education (CULT) are each associated to the legislative work under Rule 57, with shared and/or exclusive competences for specific aspects of the proposal. Parliament adopted its negotiating position (499 votes in favour, 28 against and 93 abstentions) on 14 June 2023, with substantial amendments to the Commission's text, including:

- **Definitions.** Parliament amended the definition of AI systems to align it with the definition agreed by the OECD. Furthermore, Parliament enshrines a definition of 'general purpose AI system' and 'foundation model' in EU law.
- **Prohibited practices.** Parliament substantially amended the list of AI systems prohibited in

the EU. Parliament wants to ban the use of biometric identification systems in the EU for both real-time and ex-post use (except in cases of severe crime and pre-judicial authorisation for ex-post use) and not only for real-time use, as proposed by the Commission. Furthermore, Parliament wants to ban all biometric categorisation systems using sensitive characteristics (e.g. gender, race, ethnicity, citizenship status, religion, political orientation); predictive policing systems (based on profiling, location or past criminal behaviour); emotion recognition systems (used in law enforcement, border management, workplace, and educational institutions); and AI systems using indiscriminate scraping of biometric data from social media or CCTV footage to create facial recognition databases.

- **High-risk AI systems.** While the Commission proposed to automatically categorise as high-risk all systems in certain areas or use cases, Parliament adds the additional requirement that the systems must pose a 'significant risk' to qualify as high-risk. AI systems that risk harming people's health, safety, fundamental rights or the environment would be considered as falling within high-risk areas. In addition, AI systems used to influence voters in political campaigns and AI systems used in recommender systems displayed by social media platforms, designated as very large online platforms under the Digital Services Act, would be considered high-risk systems. Furthermore, Parliament imposes on those deploying a high-risk system in the EU an obligation to carry out a fundamental rights impact assessment.
- **General-purpose AI, generative AI and foundation models.** Parliament sets a layered regulation of general-purpose AI. Parliament imposes an obligation on providers of foundation models to ensure robust protection of fundamental rights, health, safety, the environment, democracy and the rule of law. They would be required to assess and mitigate the risks their models entail, comply with some design, information and environmental requirements and register such models in an EU database. Furthermore, generative foundation AI models (such as ChatGPT) that use large language models (LLMs) to generate art, music and other content would be subject to stringent transparency obligations. Providers of such models and of generative content would have to disclose that the content was generated by AI not by humans, train and design their models to prevent generation of illegal content and publish information on the use of training data protected under copyright law. Finally, all foundation models should provide all necessary information for downstream providers to be able to comply with their obligations under the AI act.
- **Governance and enforcement.** National authorities' competences would be strengthened, as Parliament gives them the power to request access to both the trained and training models of the AI systems, including foundation models. Parliament also proposes to establish an AI Office, a new EU body to support the harmonised application of the AI act, provide guidance and coordinate joint cross border investigations. In addition, Members seek to strengthen citizens' rights to file complaints about AI systems and receive explanations of decisions based on high-risk AI systems that significantly impact their rights.
- **Research and innovation.** To support innovation, Parliament agrees that research activities and the development of free and open-source AI components would be largely exempted from compliance with the AI act rules.

Policy debate latest issues. *The recent and rapid development of general-purpose artificial intelligence technologies has framed the policy debate around, inter alia, defining general-purpose AI models, the application of the EU copyright framework to generative AI, how to ensure foundation models' compliance with AI Act principles, and the design of efficient auditing procedures for large language models (LLMs). A risk of over-regulation detrimental for investment in AI in the EU has been identified should overly stringent obligations of risk assessment, mitigation and management be imposed on foundation models and on SMEs. How to set pro-competitive rules for sandboxing and open-source AI systems has also been discussed. While there are concerns that AI poses societal-scale risks similar to nuclear weapons, calls for* *a pause in AI development have been made by civil society organisations, AI experts and tech executives. The question how to address dual-use and military AI applications has also been raised. Furthermore, given EU regulation will take time to take effect, the adoption of voluntary codes of conduct and of an AI Pact are envisaged to mitigate the potential downsides of generative AI. A pressing issue is to set a common terminology so that lawmakers around the globe have the same understanding of the technologies they need to address.*

KEY TAKEAWAYS

• **AI Regulatory Framework:** The European Commission has proposed a regulatory framework for AI, aiming to ensure AI's trustworthiness and uphold EU values. This framework is the first of its kind globally and seeks to balance innovation with safety and fundamental rights.

• **Risk-based Approach:** The draft AI act categorizes AI systems based on risk levels: unacceptable, high, limited, and low or minimal risk. The level of regulation is tailored to these risk categories, with stricter regulations for higher-risk systems.

• **Prohibited and High-risk AI Practices:** Certain AI practices, deemed to pose an 'unacceptable risk', are explicitly banned. High-risk AI systems, on the other hand, are subject to stringent rules, including conformity assessments, transparency requirements, and data governance standards.

• **Governance and Enforcement:** The proposal mandates Member States to designate competent authorities for supervision and enforcement. Non-compliance with the AI act can result in significant fines, with penalties reaching up to €30 million or 6% of the total worldwide annual turnover.

• **Stakeholder and Academic Views:** Stakeholders and academics have raised concerns about the proposed act. Major issues include the definition of 'AI systems', the effectiveness of the risk-based approach, and the need for stronger enforcement mechanisms. There's also a debate on the potential over-regulation and its impact on AI investments in the EU.

• **Legislative Process and Policy Debate:** Both the Council and Parliament have proposed amendments to the Commission's initial proposal. The rapid development of AI, especially general-purpose AI, has sparked policy debates on definitions, potential societal risks, global coordination, and the need for a common terminology.

Operationalizing AI Standards: A European Outlook

Contributors

Julien Chiaroni

Former Director of the Artificial General Secretary of Investment in the French Innovation Council

Dr. Konstantinos Karachalios

Managing Director of the IEEE Standards Association

Dr. Sebastian Hallensleben

Head of Digitalisation and AI at VDE, and Chair of CEN-CENELEC JTC 21

Operationalizing AI Standards: A European Outlook

By Julien Chiaroni, Dr. Konstantinos Karachalios, and Dr. Sebastian Hallensleben

This article is based on an EAIGG Perspectives Series Panel discussion. To access the original conversation, click here.

In a recent talk hosted by EAIGG, titled Establishing Ethical AI Standards, panelists from leading standards bodies in Europe discussed the meaning and definition of Trustworthy AI – in all its component parts – and explored models and processes of measuring these characteristics in a more granular fashion. EAIGG community members received a sneak peak into one of the largest public private partnerships in the world focused on operationalizing AI standards in accordance with the EU AI Act. The French-led initiative, Confiance.ai, brings together international standards bodies like IEEE and Germany's VDE to create a uniform approach to norms, standards, and labeling of trustworthy AI, in line with industry players in developing trustworthy applications for critical systems. The panel featured Dr. Konstantinos Karachalios, managing director of the IEEE Standards Association; Julien Chiaroni, former director of the Artificial General Secretary of Investment in the French Innovation Council; and Dr. Sebastian Hallensleben, Head of Digitalisation and AI at VDE, and Chair of CEN-CENELEC JTC 21, the European Commission's premier AI standards body. The session aimed to shed light on the collaborative efforts and strategic directions in AI ethics and governance.

Defining Trustworthy AI

There are a range of different bodies, initiatives and approaches to addressing Trustworthy AI globally, and even within Europe. The EU AI Act explores a voluntary labeling scheme aimed at enhancing protection against high-risk systems, aligning with the broader goal of ensuring safety and trust in AI applications. The IEEE 7000 series sets standards for translating ethical considerations into system requirements and design practices. In Germany, VDE has led the creation of an AI trust standard, complementing the approaches taken by France and IEEE. Despite a growing consensus and harmonization across these different initiatives, however, a clear definition of "trustworthy AI" remains elusive.

The conversation begins with a reflection on the consistency found in the wave of white papers on AI ethics that emerged a few years ago. Despite variations in terminology, principles like transparency, fairness, explainability, accountability, and robustness are remarkably consistent. The approach to defining trustworthiness involves analyzing these papers to identify the key component parts of each term, in a way that is clear, concrete and measurable. While concepts like fairness and transparency are complex, breaking them down into specific criteria makes them more manageable.

Building on these points, trustworthiness can be described as an umbrella term encompassing various features, behaviors, and properties. The idea of breaking down high-level concepts into measurable criteria allows for benchmarking and measurement of specific aspects. For example, criteria have been developed to determine whether a user is aware they are speaking to a robot and not a human. This approach makes high-level aspirations more concrete and actionable. To date, 250 criteria have been identified by these bodies to measure aspects like transparency, accountability, minimization of bias, and respect for privacy.

The conversation also emphasizes the importance of focusing on systems using AI rather than just AI itself. This perspective allows for a more comprehensive understanding of the application and its components. Acknowledging the challenge in defining trust, and specifying the criteria that make up trustworthiness, amplifies the need for collaborative work to be done between system developers and regulators in defining terminology, specifications, and KPIs.

Development of Trustworthy AI Systems

Moving to the development and design of trustworthy AI systems, panelists spotlight challenges in defining key criteria, such as explainability. The complexity lies in the varying types and contexts of AI systems, making it essential to create a clear scoring system to demonstrate conformity. The methodologies and tools used to achieve this emphasize the importance of a unified approach to trustworthiness.

A clear distinction is made, panelists argue, between the processes for designing new systems and for those assessing existing ones. While engineers have the tools to handle high-integrity, complex systems, AI's adaptive and autonomous nature requires a new consultative approach. This includes considering the societal impact of AI, especially in areas like internet platforms and technology addiction. The need for new process models that combine engineering processes with a human-centered approach is of paramount importance, they argue, calling for collaboration with other disciplines.

This also requires looking at characteristics on the product level, which can be concrete and measurable, and then assessing an organization's willingness and maturity to produce trustworthy AI. Finally, the panelists advocate for highlighting mutual mappings between frameworks rather than merging them into one unwieldy structure. This approach allows for complementarity and a more coherent framework for AI ethics and integrity.

Introducing Confiance.ai

At the forefront of government-led efforts to support trustworthy AI governance, Julien Chiaroni introduced Confiance.ai. This initiative, backed by the Prime Minister's Office, was launched four years ago to promote the development of trustworthy AI in France, Europe, and abroad, and focuses on compliance with new EU-wide regulations, future standards development, and norms adoption. Emphasizing risk management and business applications the program brings together hundreds of engineers and AI professionals across a range of industrial level use cases in the high risk category of AI deployments.

The three main pillars of Confiance.ai are 1) Standardization; 2) Conformity assessments; and 3) Engineering tools and methodologies. Chiaroni highlights cooperation with organizations like IEEE and VDE, along with investments in conformity assessments and certifications to bring products to market. This draws focus to trustworthy data engineering tools and methodologies, partnering with various industries like economics, automotive, defense, and IT. The effort seeks to develop concrete solutions for application development, certification, and linking horizontal standardization with vertical business decisions.

The program, one of the largest research and technological initiatives on trustworthy AI in Europe, is divided into seven sub-projects based on industrial use cases, covering technological aspects, design, evaluation, runtime monitoring, methodologies, and AI-based development. The ultimate aim is to transform methodologies and technologies in AI into standards that can be shared, ensuring timely standards for applications.

Use cases for Confiance.ai

The criticality of genuine industrial applications in shaping AI tools and approaches is evident. With a focus on actionable data and insights, the industrial realm's openness to offering diverse use cases is commendable, ranging from the nuances of computer vision within Industry 4.0 to the intricacies of commendable, ranging from the nuances of computer vision within Industry 4.0 to the intricacies of autonomous driving. These varied applications enhance the versatility of the tools developed. However, a challenge arises when transitioning from basic functionalities to critical system applications. This demands meticulous monitoring, dedicated embedded hardware, and sophisticated components.

Further complexities arise when attempting to exchange insights across similar industries, prompting shifts in specifications and oversight. Current endeavors emphasize system applications, Natural Language Processing (NLP), and a fusion of AI techniques, with outcomes expected by year-end. Through collective efforts and a focus on practical solutions, Confiance.ai is positioned to make significant advancements in AI technology.

Harmonized Standards Across Europe

Recognizing the urgency for harmonized AI trustworthiness standards, especially in the context of the EU AI Act, there's a palpable momentum in the European landscape to merge existing efforts. Given the restrictive timelines and the immense challenges posed by developing such standards from the ground up, the robust methodologies developed in France and Germany emerge as foundational pillars for the continental effort.

With the standardization world's rigorous procedures leaving a slender timeframe—around a year—for true development, there's a collective consensus on the necessity to build upon existing frameworks. The Franco-German alliance, in particular, plays a pivotal role in this narrative, enhancing the velocity and direction of this collaborative endeavor.

It's notable that significant standards, like the 7000 series, were spearheaded by Europeans and deeply embed European values. Public records from the European Commission provide insights into the origin and progression of these standards, highlighting the strides made. Furthermore, there's an emphasis on localizing global intelligence, adapting overarching values and guidelines to cater to specific regional and national imperatives.

Amidst this complex political landscape, the commitment is not just in the ideation but in the execution—engineers, stakeholders, and decision-makers coming together to refine and adapt the established foundations to align with regulatory visions. With partners like IEEE, a globally renowned standards body, joining hands with regional giants, the aspiration isn't just harmonized standards, but resilient systems that serve industries and societies alike. The unity in approach and diversity in expertise promises a promising trajectory toward realizing this vision.

The Future of Broader Standardization

Looking ahead, there is an expansive standardization movement in AI, spotlighting essential milestones, trajectories, and obstacles. Government regulators and industry leaders both face a palpable urgency to act, particularly in light of AI's relentless pace.

The first milestone is crafting a comprehensive set of core standards as soon as possible. While the foundational architectures are firmly positioned, the immediate journey demands rigorous consensus-building. Panelists emphasize the adoption of a modular and agile approach, innovating the way standards are created to keep pace with technological advancements.

Pioneering efforts aim to invert the conventional hierarchy of standards generation, initiating with detailed specifications and subsequently molding them into universal standards. This unorthodox strategy, adopted by Confiance.ai, helps appease the dynamic political landscape while catalyzing industry collaboration and grounded action in the private sector.

Additionally, standards bodies have seen movement towards sculpting horizontal, adaptable standards for diverse application realms. Generative AI, including deep fakes, has been identified as a significant area of concern and interest. The potential of this technology to undermine trust in the digital space has been well acknowledged, alongside louder calls for a fundamental standardization and policy response. The mainstream urgency around generative AI is seen as an opportunity to shape this area, managing both its innovative potential and its downsides. Experts are rigorously discussing how existing standards may fit, or clash, with generative AI use cases, particularly as this explosive new tool introduces widespread B2C use cases, well outside the realm of industrial and enterprise use cases discussed earlier.

In any case, the impetus is clear: European governments, alongside technology hubs around the world, face an urgency to conventional processes, and roll out standardization efforts. While the task is intricate and multi-layered, mutual cooperation is indispensable in achieving harmony and unity, and great advances are well underway to foster a comprehensive, adaptable framework championing the EU's AI Act. Navigating this labyrinth might be challenging, but the potential rewards in the forthcoming year are immense, and regulators have an appetite to move swiftly. Recognizing the depth of this initiative, spearheaded by France and supported by powerhouses like Germany and the IEEE, one can't help but wonder: Will this endeavor echo the global influence of GDPR on data regulation? Will they transform the EU into a global model that other advanced economies later mimic? Only time will tell.

KEY TAKEAWAYS

• **Defining Trustworthy AI:** The article discusses the ongoing efforts to define and measure Trustworthy AI in Europe. Despite a growing consensus, a clear definition remains elusive. The approach involves analyzing white papers on AI ethics to identify key components like transparency, fairness, explainability, accountability, and robustness, making them clear, concrete, and measurable. The article emphasizes the importance of focusing on systems using AI rather than just AI itself, highlighting the need for collaboration between system developers and regulators.

• **Development of Trustworthy AI Systems:** The development and design of trustworthy AI systems face challenges in defining key criteria such as explainability. The article highlights the need for a clear scoring system to demonstrate conformity and a unified approach to trustworthiness. It underscores the importance of considering the societal impact of AI and calls for new process models that combine engineering processes with a human-centered approach.

• **Introduction of Confiance.ai:** Confiance.ai, a French-led initiative, is at the forefront of government-led efforts to support trustworthy AI governance. It focuses on compliance with new EU-wide regulations, future standards development, and norms adoption. The three main pillars of Confiance.ai are Standardization, Conformity assessments, and Engineering tools and methodologies. The initiative seeks to develop concrete solutions for application development, certification, and linking horizontal standardization with vertical business decisions.

• **Use Cases for Confiance.ai:** The article emphasizes the criticality of genuine industrial applications in shaping AI tools and approaches. Confiance.ai is focused on practical solutions and is positioned to make significant advancements in AI technology, covering technological aspects, design, evaluation, runtime monitoring, methodologies, and AI-based development.

• **Harmonized Standards Across Europe and Future of Broader Standardization:** The article discusses the urgency for harmonized AI trustworthiness standards in the context of the EU AI Act. It highlights the role of the Franco-German alliance in enhancing the velocity and direction of this collaborative endeavor. Looking ahead, the article emphasizes the adoption of a modular and agile approach to keep pace with technological advancements. It also discusses the movement towards sculpting horizontal, adaptable standards for diverse application realms, including concerns and interests related to generative AI.

Senate Hearing on AI Oversight: IBM's Testimony on Rulemaking

Christina Montgomery

Biography

Christina is currently the Chief Privacy & Trust Officer and AI Ethics Board Chair at IBM, where she oversees the company's global privacy program, compliance, and strategy. She earned her JD from Harvard Law School and began her legal career at Seward & Kissel LLP in New York City. Over her extensive tenure at IBM, she has held pivotal roles such as Secretary to IBM's Board of Directors, Managing Attorney, and VP, Assistant General Counsel and Secretary, playing a crucial role in the company's strategic initiatives, cybersecurity, and legal support. Christina is also an active member of the IAPP AI Governance Center Advisory Board and the Women Leading Privacy Advisory Board. She contributes her expertise to the U.S. Chamber of Commerce as a Member of the Commission on Artificial Intelligence Competitiveness, Inclusion, and Innovation and serves on the National Artificial Intelligence Advisory Committee for the United States Department of Commerce.

Senate Hearing on AI Oversight: IBM’s Testimony on Rulemaking

By Christina Montgomery

Christina Montgomery shaking hands with Chairman Blumenthal before her testimony.

Chairman Blumenthal, Ranking Member Hawley, members of the Subcommittee: Thank you for today’s opportunity to present before the subcommittee. My name is Christina Montgomery, and I am IBM’s Chief Privacy and Trust Officer. I also co-chair our company’s AI Ethics Board.

Introduction

AI is not new, but it has advanced to the point where it is certainly having a moment. This new wave of generative AI tools has given people a chance to experience it first-hand. Citizens are using it for help with emails, their homework, and so much more. While IBM is not a consumer-facing company, we are just as active – and have been for years – in helping business clients use AI to make their supply chains more efficient, modernize electricity grids, and secure financial networks from fraud. IBM’s suite of AI tools, called IBM Watson after the AI system that won on TV’s Jeopardy! more than a decade ago, is widely used by enterprise customers worldwide. Just recently we announced a new set of enhancements, called watsonx, to make AI even more relevant today. Our company has extensive experience in the AI field in both an enterprise and cutting-edge research context, and we could spend an entire afternoon talking about ways the technology is being used today by business and consumers. But the technology's dramatic surge in public attention has, rightfully, raised serious questions at the heart of today's hearing. What are AI's potential impacts on society? What do we do about bias? What about misinformation, misuse, or harmful and abusive content generated by AI systems? Senators, these are the right questions, and I applaud you for convening today's hearing to address them head-on. IBM has strived for more than a century to bring powerful new technologies like artificial intelligence into the world responsibly, and with clear purpose. We follow long-held principles of trust and transparency that make clear the role of AI is to augment, not replace, human expertise and judgment. We were one of the first in our industry to establish an AI Ethics Board, which I co-chair, and whose experts work to ensure that our principles and commitments are upheld in our global business engagements.2 And we have actively worked with governments worldwide on how best to tailor their approaches to AI regulation. It's often said that innovation moves too fast for the government to keep up. But while AI may be having its moment, the moment for the government to play its proper role has not passed us by. This period of focused public attention on AI is precisely the time to define and build the right guardrails to protect people and their interests. It is my privilege to share with you IBM's recommendations for those guardrails.

Precision Regulation

The hype around AI has created understandable confusion among some in government on where intervention is needed and how regulatory guardrails should be shaped. But at its core, AI is just a tool, and tools can serve different purposes. A wrench can be used to assemble a desk or construct an airplane, yet the rules governing those two end products are not primarily based on the wrench — they are based on use. That is why IBM urges Congress to adopt a "precision regulation" approach to artificial intelligence. This means establishing rules to govern the deployment of AI in specific use-cases, not regulating the technology itself. A precision regulation approach that we feel strikes an appropriate balance between protecting Americans from potential harms and preserving an environment where innovation can flourish would involve:

- **Different Rules for Different Risks** – A chatbot that can share restaurant recommendations or draft an email has different impacts on society than a system that supports decisions on credit, housing, or employment. In precision regulation, the more stringent regulation should be applied to the use-cases with the greatest risk.
- **Clearly Defined Risks** – There must be clear guidance on AI end uses or categories of AI-supported activity that are inherently high-risk. This common definition is key to ensuring that AI developers and deployers have a clear understanding of what regulatory requirements will apply to a tool they are building for a specific end use. Risk can be assessed in part by considering the magnitude of potential harm and the likelihood of occurrence.
- **Be Transparent, Don't Hide Your AI** – Americans deserve to know when they are interacting with an AI system, so Congress should formalize disclosure requirements for certain uses of AI. Consumers should know when they are interacting with an AI system and whether they have recourse to engage with a real person, should they so desire. No person, anywhere, should be tricked into interacting with an AI system. AI developers should also be required to disclose technical information about the development and performance of an AI model, as well as the data used to train it, to give society better visibility into how these models operate. At IBM, we have adopted the use of AI Factsheets – think of them as similar to AI nutrition information labels – to help clients and partners better understand the operation and performance of the AI models we create.
- **Showing the Impact** – For higher-risk AI use-cases, companies should be required to conduct impact assessments showing how their systems perform against tests for bias and other ways that they could potentially impact the public, and attest that they have done so. Additionally,

bias testing and mitigation should be performed in a robust and transparent manner for certain high-risk AI systems, such as law enforcement usecases. These high-risk AI systems should also be continually monitored and re-tested by the entities that have deployed them.

IBM recognizes that certain AI use-cases raise particularly high levels of concern. Law enforcement investigations and credit applications are two often-cited examples. By following the risk-based, use-case specific approach at the core of precision regulation, Congress can mitigate the potential risks of AI without stifling its use in a way that dampens innovation or risks cutting Americans off from the trillions of dollars of economic activity that AI is predicted to unlock.

Generative AI

The explosion of generative AI systems in recent months has caused some to call for a deviation from a risk-based approach and instead focus on regulating AI in a vacuum, rather than its application. This would be a serious error, arbitrarily hindering innovation and limiting the benefits the technology can provide. A riskbased approach ensures that guardrails for AI apply to any application, even as this new, potentially unforeseen developments in the technology occur, and that those responsible for causing harm are held to account.

When it comes to AI, America need not choose between responsibility, innovation, and economic competitiveness. We can, and must, do all three now.

Business' Role

This focus on regulatory guardrails established by Congress does not – not by any stretch – let business off the hook for its role in enabling the responsible deployment of AI. I mentioned that IBM has strong AI governance practices and processes in place across the full scope of our global enterprise. We have principles grounded in ethics and people-centric thinking, and we have strong processes in place to bring them to life. This is also good business: IBM has long recognized ethics and trustworthiness are key to AI adoption, and that the first step in achieving these is the adoption of effective risk management practices. Companies active in developing or using AI must have (or be required to have) strong internal governance processes, including, among other things:

- Designating a lead AI ethics official responsible for an organization's trustworthy AI strategy
- Standing up an AI Ethics Board or similar function to serve as a centralized clearinghouse for resources to help guide implementation of that strategy.

IBM has taken both steps and we continue calling on our industry peers to follow suit.

Our AI Ethics Board plays a critical role in overseeing our internal AI governance process, creating reasonable internal guardrails to ensure we introduce technology into the world in a responsible and safe manner. For example, the board was central in IBM's decision to sunset our general purpose facial recognition and analysis products, considering the risk posed by the technology and the societal debate around its use. IBM's AI Ethics Board infuses the company's principles and ethical thinking into business and product decision-making. It provides centralized governance and accountability while still being flexible enough to support decentralized initiatives across IBM's global operations.

The board, along with a global community of AI Ethics focal points and advocates, reviews technology use-cases, promotes best practices, conducts internal education, and leads our participation with stakeholder groups worldwide. In short, it is a mechanism by which IBM holds our company and all IBMers accountable to our values, and our commitments to the ethical development and deployment of technology.

We do this because we recognize that society grants our license to operate. If businesses do not behave responsibly in the ways they build and use AI, customers will vote with their wallets. And with AI, the stakes are simply too high, the technology too powerful, and the potential ramifications too real. AI is not some fun experiment that should be conducted on society just to see what happens or how much innovation can be achieved.

If a company is unwilling to state its principles and build the processes and teams to live up to them, it has no business in the marketplace.

Conclusion

Mr. Chairman, and members of the subcommittee, the era of AI cannot be another era of moving fast and breaking things. But neither do we need a six-month pause – these systems are within our control today, as are the solutions. What we need at this pivotal moment is clear, reasonable policy and sound guardrails. These guardrails should be matched with meaningful steps by the business community to do their part. This should be an issue where Congress and the business community work together to get this right for the American people. It's what they expect, and what they deserve. IBM welcomes the opportunity to work with you, colleagues in Congress, and the Biden Administration to build these guardrails together. Thank you for your time, and I look forward to your questions.

KEY TAKEAWAYS

- **IBM's Role:**
  - AI's growing prevalence has raised concerns about societal impacts.
  - IBM's AI suite, IBM Watson, is widely used and has been enhanced with watsonx.
  - IBM has a long history of promoting responsible AI deployment and has established an AI Ethics Board.

- **Precision Regulation:**
  - IBM advocates for regulating AI based on specific use-cases, not the technology itself.
  - Different AI applications should have tailored regulations based on societal impact.
  - Transparency is essential, with clear disclosures when people interact with AI and details about AI's technical workings.

- **Generative AI & Business Responsibility:**
  - A risk-based approach to regulating generative AI is more effective than broad regulation.
  - Businesses, like IBM, have a significant role in ensuring the ethical deployment of AI.
  - IBM emphasizes strong AI governance, with designated ethics officials and an active AI Ethics Board.

- **Way Forward:**
  - The AI era requires careful and thoughtful progression.
  - Clear policies and guardrails are essential, with businesses playing a pivotal role.
  - Collaboration between various stakeholders, including Congress and businesses, is crucial for AI's responsible future.

Policymaking in the Pause

Future of Life Institute

Biography

The Future of Life Institute (FLI) is a nonprofit organization founded in 2014 that works to mitigate existential risks facing humanity. Based in Boston, FLI was co-founded by MIT professor Max Tegmark along with robotics expert Stuart Russell, Skype co-founder Jaan Tallinn, and entrepreneur Anthony Aguirre.

FLI focuses on reducing threats from artificial intelligence, nuclear weapons, biotechnology, and climate change. They promote research and initiatives aimed at ensuring that powerful technologies are beneficial for humanity. Some of their main programs include AI safety research grants, nuclear security fellowships, and biotechnology policy. FLI aims to support the development of new technologies while also minimizing risks. Through grants, education, and policy advocacy they strive to safeguard the future wellbeing of humans, animals, and the environment. FLI brings together researchers, policy experts, philanthropists and influential leaders to collaborate across institutions and borders to address humanity's biggest challenges.

Miriam Vogel

Biography

Miriam Vogel is the President and CEO of EqualAI, a non-profit created to reduce unconscious bias in our AI and promote responsible AI governance. Miriam co-hosts a podcast, "In AI we Trust," with the World Economic Forum and has taught Technology Law and Policy at Georgetown University Law Center, where she serves as chair of the alumni board, and also serves on the senior advisory board to the Center for Democracy and Technology (CDT). Previously, Miriam served in the U.S. government leadership, including positions in the three branches of federal government. At the Department of Justice, she served as Associate Deputy Attorney General, where she advised the Attorney General and the Deputy Attorney General (DAG) on a broad range of legal, policy, and operational issues. Miriam served in the White House in two Administrations, most recently as the Acting Director of Justice and Regulatory Affairs. Miriam previously served as General Counsel at WestExec Advisors and Associate General Counsel at Dana-Farber Cancer Institute, and practiced entertainment/ corporate transactional law at Sheppard Mullin in Los Angeles. Miriam began her legal career as a federal clerk in Denver, Colorado after graduating from Georgetown University Law Center and is a third-generation alumna from the University of Michigan.

All committee member biographies are available in the Full Report here.

National AI Advisory Committee 2023 Report

The following piece has been excerpted from the National AI Advisory Committee's first Annual Report, issued in May 2023, which is available here in full.

Introduction

Artificial intelligence (AI) can unlock significant opportunities for individuals, organizations, businesses, the economy, and society. AI can fuel life-saving advances in healthcare, enhance educational training and workforce readiness, and facilitate the equitable distribution of opportunity. AI also powers many everyday products and services, and this is only likely to increase as the applicability and usefulness of AI advances. In the last few months alone, our awareness of and interest in AI in our daily lives has increased significantly. The release of powerful new AI technologies to the general public — such as Generative AI and Large Language Models (LLMs) — has opened eyes and imaginations to the potential and versatility of AI. We have seen that AI has the potential to power and propel the American economy by enabling innovation and productivity for a broader cross section of our population. AI also has the potential to help address many of society's greatest opportunities and challenges. It can assist with scientific discovery in the health and the life sciences. It can help with climate science and sustainability. And it can help people today survive or avoid natural disasters, with innovations like wildfire and flood forecast alerts.

However, like many new technologies, AI also presents challenges and risks to both individuals and society. For example, AI systems used to attract and retain talent in the workforce can expand opportunity, but could also amplify and perpetuate historical bias and discrimination at unprecedented speed and scale. Further, AI could be misused in harmful ways, such as spreading disinformation or engaging in cybercrime. AI systems could help enhance access, such as accommodating individuals with disabilities or linguistic barriers, or it could deliver incorrect diagnoses. AI could create economic opportunity or worsen the digital divide for individuals and communities. In the workforce, we are likely to see growth of new occupations and decline of others, as well as ongoing changes to many more occupations. All such challenges magnify the need for appropriate AI oversight and safeguards.

The balance we establish in addressing these two divergent AI realities — fully harnessing its benefits while also effectively addressing its challenges and risks — will significantly impact our future. If navigated appropriately, the U.S. government can ensure that AI creates greater opportunities, providing economic and societal benefits for a broader cross section of the population. However, if navigated poorly, AI will further widen the opportunity gap, and trustworthy AI for all may become an unrealized aspiration. The importance of this moment extends beyond domestic borders, and the U.S. has an essential leadership role on the global stage in ensuring we understand and achieve trustworthy AI. The U.S. must proactively establish mandates and mechanisms to advance trustworthy AI and avoid ceding AI leadership to those entities with less equitable and inclusive goals.

The National Artificial Intelligence Advisory Committee (NAIAC) was created to advise the President on the intersection of AI and innovation, competition, societal issues, the economy, law, international relations, and other areas that can and will be impacted by AI in the near and long term. Committee members hail from diverse backgrounds — academia, industry, civil society, government — and all possess deep and complementary expertise in AI.

Here, we present our year-one findings: high-level themes, our objectives, proposed actions, and a plan for future Committee activities. Our goal is to help the U.S. government and society at large navigate this critical path to harness AI opportunities, create and model values-based innovation, and reduce AI's risks. Our findings are grounded on core beliefs, such as: the establishment of safe and effective AI systems that are opportunity-creating and beneficial to society; there must exist robust defenses against algorithmic discrimination, including support for civil rights and civil liberties; data privacy is paramount; and people deserve to know if automated decision making is being used — and should always have a recourse like human intervention.

This report is divided into four thematic AI areas, based on our focused efforts over the past year, guided by the concerns listed in our statutory mandate including: Leadership in Trustworthy Artificial Intelligence, Leadership in Research and Development, Supporting the U.S. Workforce and Providing Opportunity, and International Collaboration. Under each theme, we provide our broad objectives for U.S. leadership, and several, more granular recommended actions. The content was developed by five working groups, with each NAIAC member serving on two working groups, and ultimately presenting the consensus of the full Committee.

There are several intended audiences for this report. In line with our congressional mandate, we write this report to advise the President and the White House in navigating AI policy. We also write for the Members of Congress, to whom we are grateful for the creation of NAIAC and for their continued support for our work, and for AI innovators and policymakers more generally. Finally, as noted in our first NAIAC meeting in May 2022, we will continue to engage a broad cross section of the population that includes underrepresented communities and geographically diverse regions. We will foster a national conversation on AI governance to better understand and achieve trustworthy AI. We will do this by creating ongoing dialogues, sharing our findings, and amplifying known and new experts in this space.

It is important to note that NAIAC's undertakings are a work-in-progress that will continue over the next two years. There are issues not addressed in this first-year report that we will focus on extensively in subsequent reports, as well as in panel discussions and other mediums. We highlight some of those areas in the final section of this report.

OVERVIEW

This is the first formal NAIAC report, and covers the first year of our three-year appointment. The report is parsed into four major themes:

1. Leadership in Trustworthy Artificial Intelligence;
2. Leadership in Research and Development;
3. Supporting the U.S. Workforce and Providing Opportunity; and
4. International Cooperation.

Under each theme, the committee offers a number of objectives for engaging with AI, from the logistical (e.g., "Bolster AI leadership, coordination, and funding in the White House and across the U.S. government") to the innovative (e.g., "Create an AI research and innovation observatory"). In total, NAIAC presents 14 objectives. Because this report is intended to be actionable, objectives are tied to recommended actions. [see full report here]

1. LEADERSHIP IN TRUSTWORTHY ARTIFICIAL INTELLIGENCE

This theme underscores the imperative of establishing a robust and reliable AI governance structure in the United States. The objectives within this theme emphasize the need for cohesive AI leadership and coordination at both federal and executive levels, tailored support for small and medium-sized organizations in AI adoption, and ensuring that AI technologies are both trustworthy and beneficial in expanding societal opportunities. Collectively, the overarching aim is to position the U.S. as a global leader in responsible AI development and deployment, emphasizing inclusivity, transparency, and accountability.

Objective 1: Operationalize Trustworthy AI Governance

In January 2023, the National Institute of Standards and Technology (NIST) introduced the AI Risk Management Framework (AI RMF) as directed by Congress. Developed with comprehensive stakeholder input, this framework is now adopted in various settings, including California. It has garnered positive feedback from diverse groups, such as Congress members, civil rights bodies, and global experts. The AI RMF emphasizes that while AI has the potential to solve complex societal issues, irresponsible development and deployment can lead to harm and rights violations. The NAIAC has actively explored the multifaceted risks associated with AI, holding public meetings in states like California and North Carolina. Trustworthy AI hinges on public trust, which in turn depends on transparency, accountability, and harm mitigation. The Administration is urged to adopt a strategy that safeguards against AI risks while reaping its benefits. The AI RMF views AI risks as both technical and societal, offering a dynamic guide for AI processes to continually identify and address risks, aiming to establish and sustain trust throughout the AI lifecycle.

NAIAC recommends the White House encourage federal agencies to implement either the AI RMF, or similar processes and policies that align with the AI RMF, to address risks in all phases of the AI lifecycle effectively, with appropriate evaluation and iteration in place.

Objective 2: Bolster AI Leadership, Coordination, and Funding in the White House and Across the U.S. Government

The U.S. government needs to unify its approach to trustworthy AI to retain its global leadership position. This involves effective coordination and funding of federal agency initiatives. Trustworthy AI should involve all stakeholders, including those affected by or involved in creating accountability systems. The report suggests various structures for AI leadership within the U.S. government, each ensuring responsible AI use and governance. Funding is crucial for leadership and coordination, especially within the White House. The National AI Initiative Office (NAIIO) has a significant role in AI coordination but is understaffed, affecting its efficiency. While the National AI Initiative Act authorized over $1 billion for AI initiatives, the funds were not fully allocated, leading to gaps in policy execution. Multiple White House Offices and federal agencies have specific roles in setting U.S. tech policy, but there's redundancy and potential confusion due to a lack of coordination. A centralized White House entity, well-resourced and organized, is recommended to streamline AI strategy and coordination across all departments and offices.

The following are some actions that the federal government can take to achieve these goals:

- Empower and fill vacant AI leadership roles in the Executive Office of the President – NAIAC advises the President and OSTP to promptly fill the vacant positions of the Director of NAIIO and the U.S. Chief Technology Officer to ensure consistent AI leadership and policy execution across the executive branch.

- Fund NAIIO to fully enact their mission – NAIAC suggests that the President or Congress allocate adequate resources for NAIIO, including at least six expert full-time staff, to ensure effective AI governance and executive coordination.
- Create a new Chief Responsible AI Officer (CRAIO) – NAIAC proposes the President establish a permanent Chief Responsible AI Officer (CRAIO) with clear responsibilities for coordinating federal AI strategies, implementing trustworthy AI principles, and interacting with other AI officers across agencies.
- Establish an Emerging Technology Council (ETC) – NAIAC recommends creating an Emerging Technology Council (ETC) led by White House leaders to guide U.S. tech policy and streamline AI coordination, focusing on civil rights, economy, and security.
- Fund NIST AI work – NAIAC stresses the importance of adequately funding NIST's AI programs under NAIIA to ensure comprehensive AI standards and collaborations, given NIST's pivotal role in AI advancements and its current underfunded status.

Objective 3: Organize and Elevate AI Leadership in Federal Agencies

The U.S. government has shown commitment to trustworthy AI through various executive orders and legislation, emphasizing its importance both within and outside the federal domain. However, recent evaluations suggest that there's room for improvement in exemplifying trustworthy AI practices. For effective AI implementation, agencies need strong leadership and strategic planning. It's crucial for each department to have a clear AI strategy that outlines their objectives, promotes trustworthy AI principles, and establishes governance structures. Research highlights the need for executive support and specialized teams for successful AI integration. Currently, there's ambiguity regarding leadership roles in the government's AI initiatives. While some orders mandate agencies to designate responsible officials for AI, these roles can be assigned to junior staff without adequate authority. In contrast, other acts have clearer stipulations, like the Foundations for Evidence-Based Policymaking Act, which mandates the appointment of specific officers, backed by clear guidance from the OMB. The NAIAC recommends ensuring senior agency leadership (e.g., a Chief AI Officer) and staff at each department or agency provide clarity and transparency, while also ensuring the executive branch captures the benefits and promotes the adoption of trustworthy AI inside and outside of government. Second, the NAIAC recommends the continued implementation of existing and forthcoming congressional mandates and executive orders on AI oversight.

Objective 4: Empower Small- and Medium-Sized Organizations for Trustworthy AI Development and Use

While many entities aim for trustworthy AI, its widespread adoption is hindered by a general knowledge and skill gap, especially among small- and medium-sized organizations (SMOs) that often lack resources for dedicated AI divisions. Trustworthy AI practices vary widely across organizations, sectors, and countries, with emerging regulations lacking clear guidance. Addressing these gaps necessitates the development of practical tools, training, and guidance on a large scale, involving collaboration from diverse partners. Some nonprofits already assist SMOs in areas like data science and privacy, and uniting these organizations can amplify their collective impact. NAIAC proposes a multi-agency task force, including representatives from various sectors, to form a public-private entity focused on developing trustworthy AI methods for SMOs. This entity would offer open-source resources, engage with impacted communities, and prioritize public benefit over commercial interests, while also contributing to international AI discussions and complementing existing educational initiatives.

Objective 5: Ensure AI is trustworthy and lawful and expands opportunities

NAIAC emphasizes the U.S. government's commitment to ensuring AI is trustworthy, lawful, and free from bias. Despite progress, there's a need for more consistent implementation of AI-specific regulations, especially as AI tools can inadvertently perpetuate discrimination. Recent actions by agencies like the DOJ and EEOC highlight the potential legal pitfalls of AI in areas like hiring and housing. President Biden has expressed a strong stance against any form of discrimination, including algorithmic bias. Various departments, including the DOJ's Civil Rights Division, are working to enforce anti-discrimination laws in the context of AI, but face challenges due to technical resource gaps and the opaque nature of some AI tools. The DOJ and other agencies have the authority to compel information for investigations, but the increasing use of AI systems in various sectors has made it more challenging to address algorithmic discrimination. NAIAC urges the U.S. government to bolster civil rights agencies with resources and tools to combat algorithmic discrimination, suggesting increased DOJ funding, integrating technologists and experts for enforcement, and exploring the use of investigative demands to address potential AI-induced legal violations.

2. LEADERSHIP IN RESEARCH AND DEVELOPMENT

This theme emphasizes the United States' ambition to remain at the forefront of AI innovation by integrating societal considerations with technical advancements. It underscores the importance of a holistic approach to AI, where research not only focuses on technological prowess but also addresses the broader societal implications and benefits. The theme advocates for comprehensive measurement tools to gauge global AI progress, inclusive resources to democratize AI research, and a strong emphasis on sociotechnical research to ensure AI developments are equitable and beneficial for all.

Objective 1: Support Socio Technical Research on AI Systems

AI systems are intertwined with societal, political, economic, and cultural contexts, necessitating a sociotechnical approach to their study and deployment. This approach goes beyond just the technical properties of AI, considering its broader implications and whether it's the right solution for a given problem. Sociotechnical research methods include observational studies, inductive reasoning, capturing the perspectives of those impacted by the technology, and evaluating AI in real-world settings. Emphasizing American values like equity and fairness, this research is crucial for AI solutions that integrate well with human systems and avoid perpetuating biases. However, the U.S. government currently lacks a clear system for identifying and funding sociotechnical research in AI, and there's a pressing need to incorporate this approach into the research environment. Moreover, U.S. policy systems are lagging in understanding AI's societal impacts, highlighting the need for research that can guide policy and legislative decisions. NAIAC urges the U.S. government to invest in sociotechnical research for AI, emphasizing the integration of societal concerns with technical development through collaboration across sectors, the development of methods, and the funding promotion of AI governance research. NIST should play a pivotal role in developing tools, standards, and best practices that prioritize equitable values in AI solutions.

Objective 2: Create an AI Research and Innovation Observatory

The U.S. government is pivotal in ensuring AI's broad societal benefits. Given AI's transformative nature, the government's decisions should be rooted in current AI knowledge, its potential applications, and areas ripe for research investments. Despite the rapid evolution of AI, there's no centralized hub for gauging AI progress or conveying insights to governmental stakeholders. To maintain U.S. AI leadership, the President should focus on three core functions: measuring, analyzing, and informing. "Measuring" involves tracking AI advancements, federal AI funding, and AI's environmental impact. "Analyzing" entails converting AI progress metrics into actionable insights, leveraging data from universities, think tanks, and NGOs. "Informing" emphasizes basing AI R&D investment decisions on comprehensive data, promoting coordinated decision-making, and establishing a feedback mechanism for AI R&D data collection, ensuring continuous refinement and stakeholder engagement. NAIAC advises the White House and Congress to follow the roadmap in "Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem," emphasizing a distributed AI resource model to prevent industry power centralization and offer a genuine alternative to the existing AI framework.

Objective 3: Create a Large-Scale National AI Research Resource

AI and machine learning advancements are reshaping the workplace, potentially automating a third of work activities in 60% of jobs by 2030. As job landscapes shift, many workers are reluctant to pursue traditional education or credentialing, making them vulnerable unless employers adopt a proactive, skills-based approach to talent management. Both private and public sectors recognize the advantages of this approach, which can enhance adaptability to economic changes and promote diversity in hiring. AI can analyze vast skills data, offering insights to guide workforce decisions. However, the federal Workforce and Labor Market Information (WLMI) system, managed by the Department of Labor (DOL) and the Bureau of Labor Statistics (BLS), currently lacks real-time, granular data to provide regional insights. Following the economic challenges of COVID-19, recommendations have been made to enhance WLMI's local data accuracy and address the evolving nature of work. DOL's 2022 plan for WLMI includes pilot programs, funding for states, and a focus on the future of work. DOL has also initiated an Enterprise Data Strategy to improve data management and promote equity for workers. NAIAC urges the DOL to modernize the WLMI system with adequate funding, emphasizing data privacy and a skills-based approach, while integrating AI tools and real-time data to support workers in the changing job landscape, considering gig workers, and adopting stringent privacy standards similar to the Census Bureau.

3. SUPPORTING THE U.S. WORKFORCE AND PROVIDING OPPORTUNITY

This theme highlights the critical role of the U.S. government in preparing the nation's workforce for the rapid advancements and integrations of AI technologies. Recognizing the transformative power of AI, the theme underscores the urgency of workforce readiness, especially given the ethical and sociotechnical challenges posed by AI. The U.S. government, as the nation's largest employer, is positioned uniquely to set a national standard for interdisciplinary AI application within its workforce. However, challenges such as digital talent shortages and the inability to compete with private sector salaries highlight the need for strategic investments and reforms. The theme also touches upon the significance of international tech talent and the need for immigration reforms to ensure the U.S. remains competitive in the global AI landscape.

Objective 1: Modernize Federal Labor Market Data for the AI Era

AI and machine learning advancements are rapidly reshaping the workplace, with projections suggesting that up to 60% of jobs could have a third of their activities automated by 2030. As this transformation unfolds, employers face challenges in efficiently matching workers with emerging opportunities. Recognizing these shifts, many, including states and the U.S. government, are adopting a skills-based approach to employment, which offers agility in economic changes and promotes diversity. AI, when used responsibly, can analyze complex skills data, offering valuable insights for workforce development. However, the federal Workforce and Labor Market Information (WLMI) system, while essential, often lacks real-time, granular data. Post-COVID-19, there's a push to enhance the WLMI, with the DOL's 2022 plan emphasizing state partnerships, funding, and understanding the future of work. Concurrently, DOL is focusing on refining its data strategy to further support the nation's workers.

NAIAC urges the DOL to prioritize and fund the modernization of the WLMI system, emphasizing worker privacy and data misuse prevention, while leveraging AI tools and a skills-based approach to enhance workforce adaptability, inclusivity, and support for diverse candidates; this includes refining data collection, expanding efforts like the "Future of Workers" group, considering gig workers' roles, and adopting stringent privacy measures akin to the Census Bureau's standards.

Objective 2: Scale an AI-Capable Federal Workforce

Advances in AI, especially Generative AI tools, are rapidly integrating into society, raising ethical and sociotechnical concerns and emphasizing the need for AI workforce readiness. The U.S. government, as the nation's largest employer, should lead in the interdisciplinary application of AI. However, it faces a significant digital talent gap, with thousands of open positions requiring digital skills. This shortage is attributed to non-competitive government salaries, lack of upskilling programs, and bureaucratic hiring processes. While initiatives like 18F, the Presidential Innovation Fellows, and the U.S. Digital Service aim to address this, they fall short of meeting the large-scale talent needs essential for maintaining the U.S.'s AI competitiveness and trustworthiness.

- NAIAC advises the Administration to create a strategy for enhancing the federal workforce's AI capabilities across three pillars: fostering AI-ready civil servants, offering AI training for current employees, and creating short-term AI federal service opportunities, with the approach being replicable and encompassing disciplines beyond STEM.
- NAIAC suggests establishing a United States Digital Service Academy to train AI-skilled civil servants, especially from underrepresented groups, and forming a Digital Service Academic Compact with colleges to further expand the AI talent pipeline, with both initiatives requiring graduates to commit to several years of government service.
- NAIAC advises the U.S. government to enhance its AI capabilities by establishing AI-specific career fields, incentivizing AI training and awareness for federal employees, promoting a skills-based hiring approach, and emphasizing diversity, equity, and inclusion in AI roles within the federal workforce.
- NAIAC urges the Administration to enhance AI proficiency in government by reinforcing existing programs, emphasizing diversity, and considering the creation of a National Reserve Digital Corps for short-term AI expertise contributions.
- NAIAC advises the U.S. to ease immigration restrictions for international tech talent, noting that over half of the AI workforce and a significant portion of AI graduates are foreign-born, yet current immigration policies hinder their ability to contribute to the U.S. AI economy.

4. INTERNATIONAL COOPERATION

This last theme calls attention to the strategic importance of the U.S. collaborating with international allies and partners in the realm of AI. Recognizing the diverse approaches to AI governance and ethics globally, the theme underscores the need for the U.S. to maintain its leadership role, fostering economic growth, protecting shared values, and ensuring individual rights. The report highlights the significance of bilateral agreements, multilateral initiatives, and joint statements that set precedents for AI coordination. It also points to the potential of collaborative research and development with allies to promote free and open societies. The theme underscores the importance of expanding and deepening international alliances, supporting global AI research initiatives, and leveraging AI for global challenges like climate change.

Objective 1: Continue to Cultivate International Collaboration and Leadership on AI

AI's global leadership is viewed by many as a competition in values, with Secretary of State Antony

Blinken emphasizing the importance of technology serving democratic values. The U.S. aims to lead in AI while upholding democratic principles and maintaining its competitive edge. The federal Workforce and Labor Market Information (WLMI) system is crucial but needs modernization to provide real-time, granular data. The U.S. seeks to engage with international partners on AI policy, emphasizing both leadership and cooperation. The European Union's AI regulations, such as the GDPR and the forthcoming AI Act, highlight the global focus on AI governance. Many countries have national AI strategies, with varying emphases on governance, economic growth, and ethics. Bilateral and multilateral agreements, like the U.S.-EU Trade and Technology Council and the U.S.-India AI Initiative, foster international AI cooperation. Organizations like the OECD and GPAI promote global collaboration on AI principles and trustworthy AI development. Maintaining global AI leadership is vital for the U.S. to ensure economic growth, protect democratic values, and maintain its global position. NAIAC advises the U.S. to bolster international alliances for a strategic AI advantage, emphasizing treaties, potential AI summits, diplomatic engagements, collaborative meetings among tech leaders, and support for existing international coalitions. Additionally, they urge the Administration to fund NIST and the Department of State to globalize the AI Risk Management Framework, promoting it as a universal language for AI risk, facilitating regulatory cooperation, and aiding companies in adhering to AI regulations internationally.

Objective 2: Create a Multilateral Coalition for the Department of Commerce (NOAA) and the Department of State to Accelerate AI for Climate Efforts

AI offers a significant opportunity to address global challenges like climate change and sustainability. While current climate models are effective on a global scale, they falter at local assessments. AI innovations, such as Earth-scale digital twins, can enhance monitoring of the planet's health, bolster transportation and supply chain resilience, and mitigate risks from extreme weather and climate disasters. However, the environmental cost of deploying large-scale AI systems, particularly in terms of electricity and water consumption, is concerning. Sustainable strategies in data science and computing are essential. Collaborative efforts between the U.S., its allies, the private sector, and academia can accelerate the development of AI solutions while minimizing environmental impacts. NAIAC recommends the Department of Commerce, through the National Oceanic and Atmospheric Administration (NOAA), and together with the Department of State, should establish a U.S.-based multilateral coalition to facilitate international cooperation around AI that supports climate and sustainability efforts.

Objective 3: Expand International Cooperation on AI Diplomacy

The U.S. Department of State recognizes emerging technology, including AI, as pivotal for foreign policy and diplomacy. Secretary Blinken introduced two new divisions to address this. The first, the Bureau of Cyberspace and Digital Policy (CDP), established in April 2022, focuses on cyberspace and digital diplomacy, aiming to promote responsible online behavior, protect internet infrastructure, serve U.S. interests, and uphold democratic values. It plays a central role in addressing AI-related challenges in national security, economic prospects, and societal effects. The second, the Office of the Special Envoy for Critical and Emerging Technology, created in January 2023, provides strategic guidance on foreign policy related to transformative technologies like AI, biotechnology, and quantum information. To effectively achieve their objectives, both the CDP and the Office of the Special Envoy require additional staff, expertise from industry and academia, coordination with other State Department entities, and resources to promote American AI innovations and governance practices. The Fiscal Year 2023 Omnibus Appropriations provided funding for the Department of State's CDP and workforce technology training, but NAIAC advises that more funds are needed to effectively staff and train the CDP for global tech diplomacy.

Objective 4: Expand International Cooperation on AI R&D

AI leadership rooted in democratic values is crucial, necessitating U.S. collaboration with global allies to set AI standards that uphold open societies. A proposed Multilateral AI Research Institute (MAIRI) offers a framework for joint research, aiming to harness the strengths of allied nations and cultivate a future-oriented global AI workforce. This institute would have a primary physical location in the U.S., supplemented by virtual engagement, and would be part of a broader network of global research entities, including national labs and universities. By consolidating resources from various nations and drawing expertise from diverse sectors, MAIRI aims to amplify collective strengths in AI development.

KEY TAKEAWAYS

- **The Role of NAIAC:** The National Artificial Intelligence Advisory Committee (NAIAC) was established to advise the President on the multifaceted implications of AI, spanning innovation, competition, societal issues, the economy, and more. Comprising experts from academia, industry, civil society, and government, NAIAC's mission is to guide the U.S. government and society in harnessing AI's potential responsibly.

- **AI's Transformative Potential:** AI technologies, especially recent advancements like Generative AI and Large Language Models (LLMs), have the potential to revolutionize various sectors, from healthcare to climate science, and can significantly impact the American economy and society at large.

- **Balancing AI's Dual Realities:** AI presents a duality of significant opportunities and inherent challenges. While it has the potential to revolutionize sectors like healthcare, education, and climate science, it also poses risks such as perpetuating biases, spreading disinformation, and exacerbating the digital divide. The U.S. government's approach to AI will be pivotal in determining whether it becomes a tool for broad societal benefit or further widens opportunity gaps.

- **Leadership in Trustworthy AI:** The U.S. aims to establish a robust AI governance structure, emphasizing trustworthy and beneficial AI deployment. Key recommendations include operationalizing the AI Risk Management Framework, bolstering AI leadership and coordination within the government, and ensuring AI's lawful and trustworthy application.

- **Leadership in R&D:** The U.S. seeks to remain at the forefront of AI innovation, integrating societal considerations with technical advancements. This involves supporting sociotechnical research, creating an AI Research and Innovation Observatory, and establishing a large-scale national AI research resource.

- **Supporting the U.S. Workforce:** AI's integration into the workforce presents both opportunities and challenges. The government aims to modernize federal labor market data for the AI era, scale an AI-capable federal workforce, and adopt a skills-based approach to employment.

- **International Cooperation:** Recognizing the global nature of AI, the U.S. emphasizes international collaboration. This includes cultivating international AI leadership, accelerating AI for climate efforts, expanding AI diplomacy, and fostering international AI R&D cooperation.

- **Future Endeavors:** NAIAC's work is ongoing, with a commitment to address more AI-related issues in subsequent reports and engagements. The committee aims to foster a national conversation on AI governance, ensuring a broad and inclusive dialogue on achieving trustworthy AI.

Algorithms Were Supposed to Reduce Bias in Criminal Justice – Do They?

Ngozi Okidegbe

Biography

Ngozi Okidegbe is an Associate Professor of Law and Assistant Professor of Computing & Data Sciences. Her focus is in the areas of law and technology, evidence, criminal procedure, and racial justice. Her work examines how the use of predictive technologies in the criminal justice system impacts racially marginalized communities. Professor Okidegbe is a Faculty Associate at the Berkman Klein Center for Internet & Society at Harvard University and an Affiliated Fellow at Information Society Project at Yale Law School. She is also on the program committee of the Privacy Law Scholars' Conference and serves on the advisory board for the Electronic Privacy Information Center. Prior to joining BU, Professor Okidegbe was an Assistant Professor at Cardozo School of Law. Before joining Cardozo, she served as a law clerk for Justice Mbuyiseli Madlanga of the Constitutional Court of South Africa and for the Justices of the Court of Appeal for Ontario. She also practiced at CaleyWray, a labor law boutique in Toronto. Professor Okidegbe's articles have been published or are forthcoming in the Critical Analysis of Law, Connecticut Law Review, UCLA Law Review, Cornell Law Review, and Michigan Law Review.

Algorithms Were Supposed to Reduce Bias in Criminal Justice – Do They?

Article originally appeared in The Brink, from Boston University

Data can discriminate, says BU's Ngozi Okidegbe, the first dual-appointed professor to the School of Law and the Faculty of Computing & Data Sciences

Algorithms were supposed to remake the American justice system. Championed as dispassionate, computer-driven calculations about risk, crime, and recidivism, their deployment in everything from policing to bail and sentencing to parole was meant to smooth out what are often unequal decisions made by fallible, biased humans.

But, so far, this hasn't been the case.

"In theory, if the predictive algorithm is less biased than the decision-maker, that should lead to less incarceration of Black and Indigenous and other politically marginalized people. But algorithms can discriminate," says Ngozi Okidegbe, Boston University's Moorman-Simon Interdisciplinary Career Development Associate Professor of Law and an assistant professor of computing and data sciences. She's the first at the University to hold a dual appointment straddling data and the law, and her scholarship dives into this intersection, examining how the use of predictive technologies in the criminal justice system impacts racially marginalized communities.

As it is, these groups are incarcerated at nearly four times the rate of their white peers. According to the Bureau of Justice Statistics, an arm of the US Department of Justice, there were 1,186 Black adults incarcerated in state or federal facilities for every 100,000 adults in 2021 (the most recent year for which data are available), and 1,004 American Indians and Alaska Natives incarcerated for every 100,000 adults. Compare these to the rates at which white people were incarcerated in the same year: 222 per 100,000.

In recent papers, Okidegbe has studied the role of algorithms in these inequities and the interwoven consequences of technology and the law, including researching the data behind bail decisions.

"Ngozi's joint appointment at the BU School of Law and in the Faculty of Computing & Data Sciences could not be more timely, as it speaks to the importance of examining and scrutinizing today's sociotechnical and human-in-the-loop AI systems and technologies," says Azer Bestavros, BU associate provost for computing and data sciences. "This scrutiny allows us not only to reimagine the design and deployment of these systems, but also to reconsider the ethical, legal, and public policy frameworks within which these systems will operate."

Algorithms Amplifying Bias

In their most basic form, algorithms are problem-solving shortcuts. Engineers can train computers to digest a large amount of data and then produce a simple solution to a complex problem. Spotify, for example, uses algorithms to suggest songs the company thinks its listeners might enjoy, based on what they've listened to previously. The more data a computer model has to go on, the more nuanced and accurate its results should be.

But a growing body of academic research—including by Okidegbe—and news reports show that algorithms built upon incomplete or biased data can replicate or even amplify that bias when they spit out results. This isn't a huge deal if, for example, your toddler's Peppa Pig obsession leaks into your suggested Spotify playlists, but it can have devastating effects in other contexts.

Consider a judge, says Okidegbe, who receives an algorithmically generated recidivism risk score as part of a report on a convicted criminal. This score tells the judge how likely this person is to commit another crime in the near future—the higher the score, the more likely someone is to be a repeat offender. The judge takes this score into account, and assigns more jail time to someone with a high recidivism score. Case closed.

A sprawling report by the nonprofit news organization ProPublica found that because these scores feel impartial, they can carry a lot of weight with the judges who use them. In reality, these scores are neither impartial nor airtight. ProPublica found that one particular system used by courts across the country guessed wrong about two times as often for Black people than for white people: it mislabeled twice as many Black people who didn't reoffend as being at high risk for doing so.

In a recent article for the Connecticut Law Review, Okidegbe traces this inconsistency back to its source, and identifies a three-pronged "input problem."

First, she writes, jurisdictions are opaque about whether and how they use pretrial algorithms, and often adopt them without consulting marginalized communities, "even though these communities are disproportionately affected by their utilization." Second, these same communities are generally shut out of the process for building such algorithms. Finally, even in jurisdictions where members of the public can lodge opinions about the use of such tools, their input rarely changes anything.

From a racial-justice perspective, there are other harms that come out of the use of these algorithmic systems. The very paradigm that governs if and how we use these algorithms is quite technocratic and not very diverse.
Ngozi Okidegbe

"From a racial-justice perspective, there are other harms that come out of the use of these algorithmic systems. The very paradigm that governs if and how we use these algorithms is quite technocratic and not very diverse. Kate Crawford has noted AI's 'white guy problem,'" Okidegbe says, referring to a principal researcher at Microsoft and co chair of a White House symposium on AI and society who coined the term to describe the overrepresentation of white men in the creation of artificially intelligent products and companies.

From the very outset, Okidegbe says, algorithmic systems exclude racially marginalized and other politically oppressed groups.

"I've been looking at the decision-making power of whether and how to use algorithms, and what data they are used to produce. It is very exclusionary of the marginalized communities that are most likely to be affected by it, because those communities are not centered, and often they're not even at the table when these decisions are being made," she says. "That's one way I suggest that the turn to algorithms is inconsistent with a racial justice project, because of the way in which they maintain the marginalization of these same communities."

Shifting Power

In addition to producing biased results that disproportionately harm marginalized communities, the data used to train algorithms can be messy, subjective, and discriminatory, Okidegbe says.

“In my work, I’ve contended with what I think is a misconception: that algorithms are only built with quantitative data. They’re not, they’re also built with qualitative data,” she says. Computer engineers and data designers will meet with policymakers to figure out what problem their algorithm should solve, and which datasets they should pull from to build it, Okidegbe says.

In the criminal and legal context, this might mean working with judges to determine what would help them deliver prison sentences, for example. Once again though, it’s much less likely that data engineers would meet with incarcerated people, say, as part of their early information-gathering process. Instead, as Okidegbe writes in an article for a recent edition of the Cornell Law Review, most large datasets used in pretrial algorithms are built upon and trained on data from “carceral knowledge sources,” such as police records and court documents.

“That puts forth this narrative that these communities have no knowledge to add toward the broader question,” Okidegbe says.

Really delivering on the promise of algorithms in the criminal justice system—the promise that they make the process more uniform and less biased than humans otherwise have—requires a radical rethinking of the entire structure, Okidegbe says. It’s something she encourages her students to consider as they shape the future of law and criminal justice.

“It means actually accounting for the knowledge from marginalized and politically oppressed communities, and having it inform how the algorithm is constructed. It also means ongoing oversight of algorithmic technologies by these communities, as well. What I am contending requires building new institutional structures, it requires shifting our mindset about who is credible and who should be in power when it comes to the use of these algorithms. And, if that is too much, then we can’t, in the same breath, call this a racial justice project.”

KEY TAKEAWAYS

- **Promise vs Reality:** Algorithms were introduced in the American justice system with the promise of eliminating human biases and making decisions based on data. However, they have not lived up to this expectation. Instead of reducing biases, in many cases, they have perpetuated or even amplified them.

- **Data Discrimination:** Ngozi Okidegbe, a dual-appointed professor at Boston University, emphasizes that while algorithms can be less biased than human decision-makers, they can still discriminate. This is especially evident in the incarceration rates of Black, Indigenous, and other politically marginalized groups, which remain disproportionately high.

- **Algorithmic Amplification of Bias:** Algorithms, when built on incomplete or biased data, can replicate or even amplify that bias. For instance, recidivism risk scores, which predict the likelihood of a person committing another crime, can be biased. A report by ProPublica found significant discrepancies in the accuracy of these scores between Black and white individuals.

- **Exclusion of Marginalized Communities:** There is a lack of transparency and inclusivity in the creation and deployment of these algorithms. Marginalized communities are often excluded from the decision-making processes related to the use of algorithms, even though they are the most affected by their outcomes.

- **Rethinking Algorithmic Implementation:** To truly harness the potential of algorithms in the criminal justice system, there needs to be a radical reimagining of their use. This includes incorporating knowledge from marginalized communities in the algorithm's construction and ensuring their ongoing oversight. The current approach to algorithms does not align with a racial justice project due to the continued marginalization of certain communities.

Policymaking in the Pause

By Future of Life Institute

Introduction

Prominent AI researchers have identified a range of dangers that may arise from the present and future generations of advanced AI systems if they are left unchecked. AI systems are already capable of creating misinformation and authentic-looking fakes that degrade the shared factual foundations of society and inflame political tensions.1 AI systems already show a tendency toward amplifying entrenched discrimination and biases, further marginalizing disadvantaged communities and diverse viewpoints.2 The current, frantic rate of development will worsen these problems significantly.

As these types of systems become more sophisticated, they could destabilize labor markets and political institutions, and lead to the concentration of enormous power in the hands of a small number of unelected corporations. Advanced AI systems could also threaten national security, e.g., by facilitating the inexpensive development of chemical, biological, and cyber weapons by non-state groups. The systems could themselves pursue goals, either human- or self-assigned, in ways that place negligible value on human rights, human safety, or, in the most harrowing scenarios, human existence.

In an effort to stave off these outcomes, the Future of Life Institute (FLI), joined by over 20,000 leading AI researchers, professors, CEOs, engineers, students, and others on the frontline of AI progress, called for a pause of at least six months on the riskiest and most resource intensive AI experiments – those experiments seeking to further scale up the size and general capabilities of the most powerful systems developed to date.

The proposed pause provides time to better understand these systems, to reflect on their ethical, social, and safety implications, and to ensure that AI is developed and used in a responsible manner. The unchecked competitive dynamics in the AI industry incentivize aggressive development at the expense of caution5 . In contrast to the breakneck pace of development, however, the levers of governance are generally slow and deliberate. A pause on the production of even more powerful AI systems would thus provide an important opportunity for the instruments of governance to catch up with the rapid evolution of the field.

We have called on AI labs to institute a development pause until they have protocols in place to ensure that their systems are safe beyond a reasonable doubt, for individuals, communities, and society. Regardless of whether the labs will heed our call, this policy brief provides policymakers with concrete recommendations for how governments can manage AI risks.

The recommendations are by no means exhaustive: the project of AI governance is perennial and will extend far beyond any pause. Nonetheless, implementing these recommendations, which largely reflect a broader consensus among AI policy experts, will establish a strong governance foundation for AI.

Policy recommendations:
1. Mandate robust third-party auditing and certification.
2. Regulate access to computational power.

3. Establish capable AI agencies at the national level.
4. Establish liability for AI-caused harms.
5. Introduce measures to prevent and track AI model leaks.
6. Expand technical AI safety research funding.
7. Develop standards for identifying and managing AI-generated content and recommendations.

To coordinate, collaborate, or inquire regarding the recommendations herein, please contact us at policy@futureoflife.org.

1. Mandate Robust Third-Party Auditing and Certification for Specific AI Systems

For some types of AI systems, the potential to impact the physical, mental, and financial wellbeing of individuals, communities, and society is readily apparent. For example, a credit scoring system could discriminate against certain ethnic groups. For other systems – in particular general-purpose AI systems6 – the applications and potential risks are often not immediately evident. General-purpose AI systems trained on massive datasets also have unexpected (and often unknown) emergent capabilities.

In Europe, the draft AI Act already requires that, prior to deployment and upon any substantial modification, 'high-risk' AI systems undergo 'conformity assessments' in order to certify compliance with specified harmonized standards or other common specifications.8 In some cases, the Act requires such assessments to be carried out by independent third-parties to avoid conflicts of interest.

In contrast, the United States has thus far established only a general, voluntary framework for AI risk assessment.9 The National Institute of Standards and Technology (NIST), in coordination with various stakeholders, is developing so-called 'profiles' that will provide specific risk assessment and mitigation guidance for certain types of AI systems, but this framework still allows organizations to simply 'accept' the risks that they create for society instead of addressing them. In other words, the United States does not require any third-party risk assessment or risk mitigation measures before a powerful AI system can be deployed at scale.

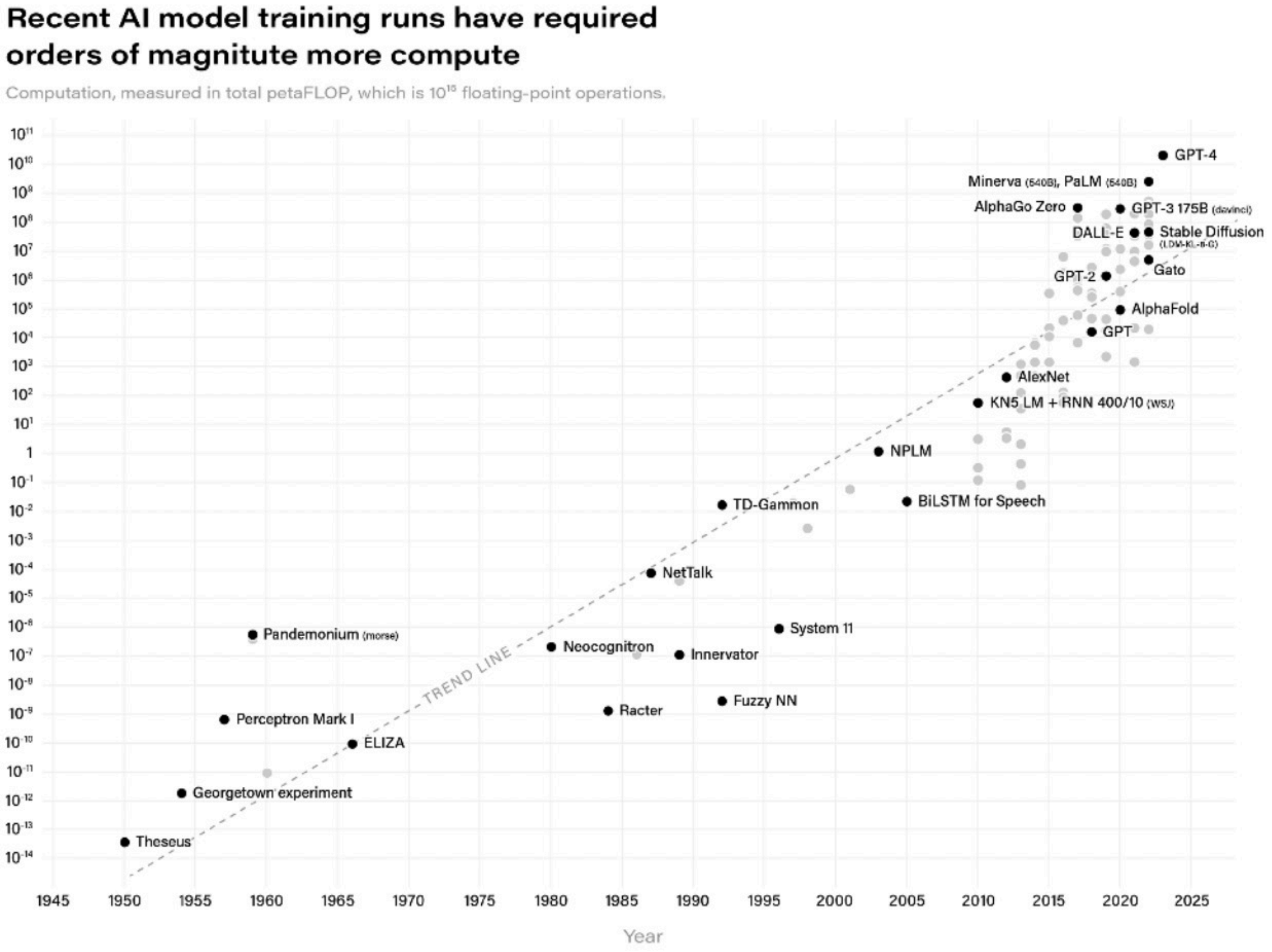
To ensure proper vetting of powerful AI systems before deployment, we recommend a robust independent auditing regime for models that are general-purpose, trained on large amounts of compute, or intended for use in circumstances likely to impact the rights or the wellbeing of individuals, communities, or society. This mandatory third-party auditing and certification scheme could be derived from the EU's proposed 'conformity assessments' and should be adopted by jurisdictions worldwide.

In particular, we recommend third-party auditing of such systems across a range of benchmarks for the assessment of risks11, including possible weaponization12 and unethical behaviors13 and mandatory certification by accredited third-party auditors before these high-risk systems can be deployed. Certification should only be granted if the developer of the system can demonstrate that appropriate measures have been taken to mitigate risk, and that any residual risks deemed tolerable are disclosed and are subject to established protocols for minimizing harm.

2. Regulate Organizations' Access to Computational Power

At present, the most advanced AI systems are developed through training that requires an enormous amount of computational power - 'compute' for short. The amount of compute used to train a general-purpose system largely correlates with its capabilities, as well as the magnitude of its risks.

Today's most advanced models, like OpenAI's GPT-4 or Google's PaLM, can only be trained with thousands of specialized chips running over a period of months. While chip innovation and better algorithms will reduce the resources required in the future, training the most powerful AI systems will likely remain prohibitively expensive to all but the best-resourced players.



Recent AI model training runs have required orders of magnitute more compute

Computation, measured in total petaFLOP, which is 10^{15} floating-point operations.

Figure 1. OpenAI is estimated to have used approximately 700% more compute to train GPT-4 than the next closest model (Minerva, DeepMind), and 7,000% more compute than to train GPT-3 (Davinci). Depicted above is an estimate of compute used to train GPT-4 calculated by Ben Cottier at Epoch, as official training compute details for GPT-4 have not been released. Data from: Sevilla et al., 'Parameter, Compute and Data Trends in Machine Learning,' 2021 [upd. Apr. 1, 2023].

In practical terms, compute is more easily monitored and governed than other AI inputs, such as talent, data, or algorithms. It can be measured relatively easily and the supply chain for advanced AI systems is highly centralized, which means governments can leverage such measures in order to limit the harms of large-scale models.

To prevent reckless training of the highest risk models, we recommend that governments make access to large amounts of specialized computational power for AI conditional upon the completion of a comprehensive risk assessment. The risk assessment should include a detailed plan for minimizing risks to individuals, communities, and society, consider downstream risks in the value chain, and ensure that the AI labs conduct diligent know-your customer checks.

Successful implementation of this recommendation will require governments to monitor the use of compute at data centers within their respective jurisdictions. The supply chains for AI chips and other key components for high-performance computing will also need to be regulated such that chip firmware can alert regulators to unauthorized large training runs of advanced AI systems.

In 2022, the U.S. Department of Commerce's Bureau of Industry and Security instituted licensing requirements[17] for export of many of these components in an effort to monitor and control their global distribution. However, licensing is only required when exporting to certain destinations, limiting the capacity to monitor aggregation of equipment for unauthorized large training runs within the United States and outside the scope of export restrictions. Companies within the specified destinations have also successfully skirted monitoring by training AI systems using compute leased from cloud providers.[18] We recommend expansion of know-your-customer requirements to all high-volume suppliers for high-performance computing components, as well as providers that permit access to large amounts of cloud compute.

3. Establish Capable AI Agencies at National Level

AI is developing at a breakneck pace and governments need to catch up. The establishment of AI regulatory agencies helps to consolidate expertise and reduces the risk of a patchwork approach.

The UK has already established an Office for Artificial Intelligence and the EU is currently legislating for an AI Board. Similarly, in the US, Representative Ted Lieu has announced legislation to create a non-partisan AI Commission with the aim of establishing a regulatory agency. These efforts need to be sped up, taken up around the world and, eventually, coordinated within a dedicated international body.

We recommend that national AI agencies be established in line with a blueprint[19] developed by Anton Korinek at Brookings. Korinek proposes that an AI agency have the power to:

- Monitor public developments in AI progress and define a threshold for which types of advanced AI systems fall under the regulatory oversight of the agency (e.g. systems above a certain level of compute or that affect a particularly large group of people).
- Mandate impact assessments of AI systems on various stakeholders, define reporting requirements for advanced AI companies and audit the impact on people's rights, wellbeing, and society at large. For example, in systems used for biomedical research, auditors would be asked to evaluate the potential for these systems to create new pathogens.
- Establish enforcement authority to act upon risks identified in impact assessments and to prevent abuse of AI systems.
- Publish generalized lessons from the impact assessments such that consumers, workers and other AI developers know what problems to look out for. This transparency will also allow academics to study trends and propose solutions to common problems.

Beyond this blueprint, we also recommend that national agencies around the world mandate record-keeping of AI safety incidents, such as when a facial recognition system causes the arrest of an innocent person. Examples include the non-profit AI Incident Database and the forthcoming EU AI Database created under the European AI Act.

4. Establish Liability for AI-Caused Harm

AI systems present a unique challenge in assigning liability. In contrast to typical commercial products or traditional software, AI systems can perform in ways that are not well understood by their developers, can learn and adapt after they are sold and are likely to be applied in unforeseen contexts. The ability for AI systems to interact with and learn from other AI systems is expected to expedite the emergence of unanticipated behaviors and capabilities, especially as the AI ecosystem becomes more expansive and interconnected.

Several plug-ins have already been developed that allow AI systems like ChatGPT to perform tasks through other online services (e.g. ordering food delivery, booking travel, making reservations), broadening the range of potential real-world harms that can result from their use and further complicating the assignment of liability. OpenAI's GPT-4 system card references an instance of the system explicitly deceiving a human into bypassing a CAPTCHA bot detection system using TaskRabbit, a service for soliciting freelance labor.

When such systems make consequential decisions or perform tasks that cause harm, assigning responsibility for that harm is a complex legal challenge. Is the harmful decision the fault of the AI developer, deployer, owner, end-user, or the AI system itself?

Key among measures to better incentivize responsible AI development is a coherent liability framework that allows those who develop and deploy these systems to be held responsible for resulting harms. Such a proposal should impose a financial cost for failing to exercise necessary diligence in identifying and mitigating risks, shifting profit incentives away from reckless empowerment of poorly-understood systems toward emphasizing the safety and wellbeing of individuals, communities, and society as a whole.

To provide the necessary financial incentives for profit-driven AI developers to exercise abundant caution, we recommend the urgent adoption of a framework for liability for AIderived harms. At a minimum, this framework should hold developers of general-purpose AI systems and AI systems likely to be deployed for critical functions strictly liable for resulting harms to individuals, property, communities, and society. It should also allow for joint and several liability for developers and downstream deployers when deployment of an AI system that was explicitly or implicitly authorized by the developer results in harm.

5. Introduce Measures to Prevent and Track AI Model Leaks

Commercial actors may not have sufficient incentives to protect their models, and their cyberdefense measures can often be insufficient. In early March 2023, Meta demonstrated that this is not a theoretical concern, when their model known as LLaMa was leaked to the internet.As of the date of this publication, Meta has been unable to determine who leaked the model. This lab leak allowed anyone to copy the model and represented the first time that a major tech firm's restricted-access large language model was released to the public.

Watermarking of AI models provides effective protection against stealing, illegitimate redistribution and unauthorized application, because this practice enables legal action against identifiable leakers. Many digital media are already protected by watermarking - for example through the embedding of company logos in images or videos. A similar process25 can be applied to advanced AI models, either by inserting information directly into the model parameters or by training it on specific trigger data.

We recommend that governments mandate watermarking for AI models, which will make it easier for AI developers to take action against illegitimate distribution.

6. Expand Technical AI Safety Research Funding

The private sector under-invests in research that ensures that AI systems are safe and secure. Despite nearly USD 100 billion of private investment in AI in 2022 alone, it is estimated that only about 100 full-time researchers worldwide are specifically working to ensure AI is safe and properly aligned with human values and intentions.

In recent months, companies developing the most powerful AI systems have either downsized or entirely abolished their respective 'responsible AI' teams. While this partly reflects a broader trend of mass layoffs across the technology sector, it nonetheless reveals the relative deprioritization of safety and ethics considerations in the race to put new systems on the market.

Governments have also invested in AI safety and ethics research, but these investments have primarily focused on narrow applications rather than on the impact of more general AI systems like those that have recently been released by the private sector. The US National Science Foundation (NSF), for example, has established 'AI Research Institutes' across a broad range of disciplines. However, none of these institutes are specifically working on the large-scale, societal, or aggregate risks presented by powerful AI systems.

To ensure that our capacity to control AI systems keeps pace with the growing risk that they pose, we recommend a significant increase in public funding for technical AI safety research in the following research domains:

- Alignment: development of technical mechanisms for ensuring AI systems learn and perform in accordance with intended expectations, intentions, and values.
- Robustness and assurance: design features to ensure that AI systems responsible for critical functions can perform reliably in unexpected circumstances, and that their performance can be evaluated by their operators.
- Explainability and interpretability: develop mechanisms for opaque models to report the internal logic used to produce output or make decisions in understandable ways. More explainable and interpretable AI systems facilitate better evaluations of whether output can be trusted.

In the past few months, experts such as the former Special Advisor to the UK Prime Minister on Science and Technology James W. Phillips and a Congressionally-established US taskforce have called for the creation of national AI labs as 'a shared research infrastructure that would provide AI researchers and students with significantly expanded access to computational resources, high-quality data, educational tools, and user support.'30 Should governments move forward with this concept, we propose that at least 25% of resources made available through these labs be explicitly allocated to technical AI safety projects.

7. Develop Standards for Identifying and Managing AI-Generated Content and Recommendations

The need to distinguish real from synthetic media and factual content from 'hallucinations' is essential for maintaining the shared factual foundations underpinning social cohesion. Advances in generative AI have made it more difficult to distinguish between AI-generated media and real images, audio, and video recordings. Already we have seen AI-generated voice technology used in financial scams.

Creators of the most powerful AI systems have acknowledged that these systems can produce convincing textual responses that rely on completely fabricated or out-of-context information.32 For society to absorb these new technologies, we will need effective tools that allow the public to evaluate the authenticity and veracity of the content they consume.

We recommend increased funding for research into techniques, and development of standards, for digital content provenance. This research, and its associated standards, should ensure that a reasonable person can determine whether content published online is of synthetic or natural origin, and whether the content has been digitally modified, in a manner that protects the privacy and expressive rights of its creator.

We also recommend the expansion of ‘bot-or-not’ laws that require disclosure when a person is interacting with a chatbot. These laws help prevent users from being deceived or manipulated by AI systems impersonating humans, and facilitate contextualizing the source of the information. The draft EU AI Act requires that AI systems be designed such that users are informed they are interacting with an AI system,[33] and the US State of California enacted a similar bot disclosure law in 2019.[34] Almost all of the world’s nations, through the adoption of a UNESCO agreement on the ethics of AI, have recognized[35] ‘the right of users to easily identify whether they are interacting with a living being, or with an AI system imitating human or animal characteristics.’ We recommend that all governments convert this agreement into hard law to avoid fraudulent representations of natural personhood by AI from outside regulated jurisdictions.

Even if a user knows they are interacting with an AI system, they may not know when that system is prioritizing the interests of the developer or deployer over the user. These systems may appear to be acting in the user’s interest, but could be designed or employed to serve other functions. For instance, the developer of a general-purpose AI system could be financially incentivized to design the system such that when asked about a product, it preferentially recommends a certain brand, when asked to book a flight, it subtly prefers a certain airline, when asked for news, it provides only media advocating specific viewpoints, and when asked for medical advice, it prioritizes diagnoses that are treated with more profitable pharmaceutical drugs. These preferences could in many cases come at the expense of the end user’s mental, physical, or financial well-being.

Many jurisdictions require that sponsored content be clearly labeled, but because the provenance of output from complex general-purpose AI systems is remarkably opaque, these laws may not apply. We therefore recommend, at a minimum, that conflict-of-interest trade-offs should be clearly communicated to end users along with any affected output; ideally, laws and industry standards should be implemented that require AI systems to be designed and deployed with a duty to prioritize the best interests of the end user.

Finally, we recommend the establishment of laws and industry standards clarifying and the fulfillment of ‘duty of loyalty’ and ‘duty of care’ when AI is used in the place of or in assistance to a human fiduciary. In some circumstances – for instance, financial advice and legal counsel – human actors are legally obligated to act in the best interest of their clients and to exercise due care to minimize harmful outcomes. AI systems are increasingly being deployed to advise on these types of decisions or to make them (e.g. trading stocks) independent of human input. Laws and standards towards this end should require that if an AI system is to contribute to the decision-making of a fiduciary, the fiduciary must be able to demonstrate beyond a reasonable doubt that the AI system will observe duties of loyalty and care comparable to their human counterparts. Otherwise, any breach of these fiduciary responsibilities should be attributed to the human fiduciary employing the AI system.

Conclusion

The new generation of advanced AI systems is unique in that it presents significant, well documented risks, but can also manifest high-risk capabilities and biases that are not immediately apparent. In other words, these systems may perform in ways that their developers had not anticipated or malfunction when placed in a different context. Without appropriate safeguards, these risks are likely to result in substantial harm, in both the near- and longer term, to individuals, communities, and society.

Historically, governments have taken critical action to mitigate risks when confronted with emerging technology that, if mismanaged, could cause significant harm. Nations around the world have employed both hard regulation and international consensus to ban the use and development of biological weapons, pause human genetic engineering, and establish robust government oversight for introducing new drugs to the market. All of these efforts required swift action to slow the pace of development, at least temporarily, and to create institutions that could realize effective governance appropriate to the technology. Humankind is much safer as a result.

We believe that approaches to advancement in AI R&D that preserve safety and benefit society are possible, but require decisive, immediate action by policymakers, lest the pace of technological evolution exceed the pace of cautious oversight. A pause in development at the frontiers of AI is necessary to mobilize the instruments of public policy toward commonsense risk mitigation. We acknowledge that the recommendations in this brief may not be fully achievable within a six month window, but such a pause would hold the moving target still and allow policymakers time to implement the foundations of good AI governance.

The path forward will require coordinated efforts by civil society, governments, academia, industry, and the public. If this can be achieved, we envision a flourishing future where responsibly developed AI can be utilized for the good of all humanity.

KEY TAKEAWAYS

• There are significant risks from advanced AI systems, including misinformation, discrimination, labor disruption, concentration of power, threats to national security, and existential threats. A temporary pause on development of the most powerful systems would allow time to implement governance measures.

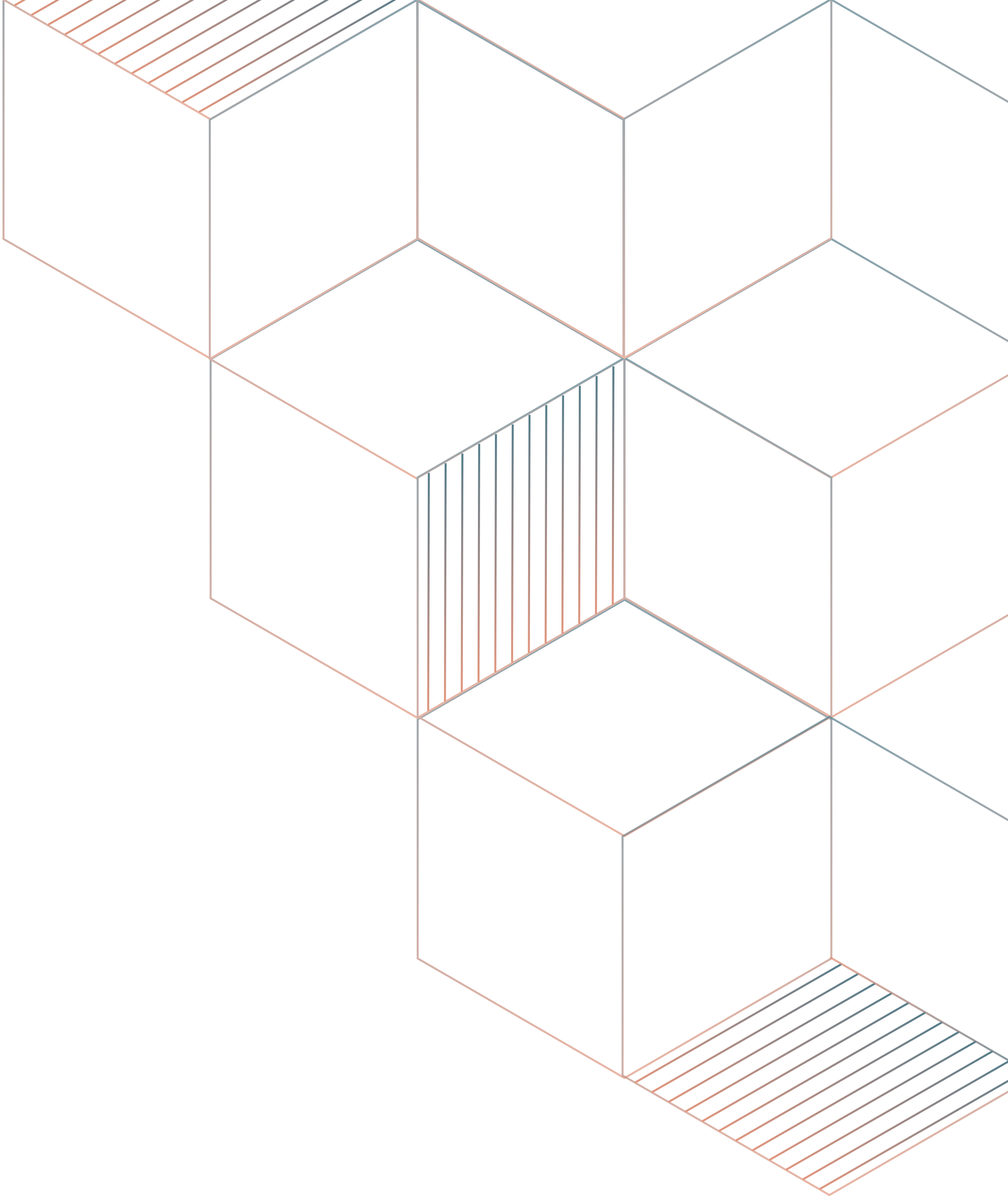
• Governments should mandate third-party auditing and certification for high-risk AI systems before deployment. Access to computational power for training AI should be regulated.

• National AI regulatory agencies should be established to oversee impacts, enforce standards, and mandate transparency. Laws are needed to assign liability for harms caused by AI systems.

• Measures should be taken to prevent and track AI model leaks, including mandated watermarking. More funding is needed for technical research into AI safety and alignment.

• Standards are needed for disclosing AI-generated content and managing conflicts of interest and fiduciary duties when AI systems are advising or acting on behalf of humans.

• Decisive action is required now to implement oversight before technological progress outpaces responsible governance of AI. With coordinated efforts, AI can be developed safely for the benefit of humanity.



INNOVATION AT THE FOREFRONT: UNLEASHING AI’S FULL POTENTIAL

Innovation Ecosystems: Benchmarking AI Disruption

Emmanuel Benhamou

Biography

Emmanuel Benhamou is the managing director for the Ethical AI Governance Group (EAIGG), and has over 20 years experience driving innovation across ethical AI, blockchain, and enterprise technology domains. He works with venture capital firms in Silicon Valley, New York, Tel Aviv and Southeast Asia, driving capital raising efforts, strategic communications campaigns and policy initiatives. Emmanuel has a deep passion for decentralized data platforms and cryptocurrency networks, and serves as a venture partner with masterkey, a venture capital firm powering the blockchain industry's foundational technology layers. Emmanuel has worked with the US Department of Defense, the US Chamber of Commerce and various other organizations that lie at the intersection of technology, security and public policy. He earned his Master of Science in Foreign Service from Georgetown University after completing his Bachelor of Science in Communications from Bosston University. He also attended the IDC's Adelson School of Entrepreneurship in Israel where he completed an executive program in Enterpreneurship & Venture Creation.

Innovation Ecosystems: Benchmarking AI Disruption

Ash Tutika

Biography

Ash has over 8 years of experience in the Venture Capital and Innovation industry where has operated, invested in and supported startups. Recently, he served as the Chief of Staff to the CEO at Arena. Before joining Arena, he worked as an Innovation Manager at One Valley, where he was responsible for overseeing products and programs aimed at supporting, investing in, and educating global entrepreneurs. Ash's experience also includes his role as an Innovation Manager at Plug and Play Ventures, where he played a key role in fostering startup investments and facilitating product development in collaboration with leading Fortune 500 companies in the Insurance and Healthcare sectors. Ash boasts a strong entrepreneurial background as well, having served as the Co-Founder and CEO of BioVirtua, a venture capital-backed healthcare analytics startup. He is a graduate of the Startupbootcamp and NASA iTech programs. Additionally, Ash actively supports the Ethical AI Governance Group, a non-profit organization comprised of AI executives, investors, and technologists working to promote responsible AI practices through open-source initiatives. Ash is currently pursuing a Master's degree in Technology Management from Georgetown University and holds a Bachelor of Science in Management Information Systems from the University of San Francisco.

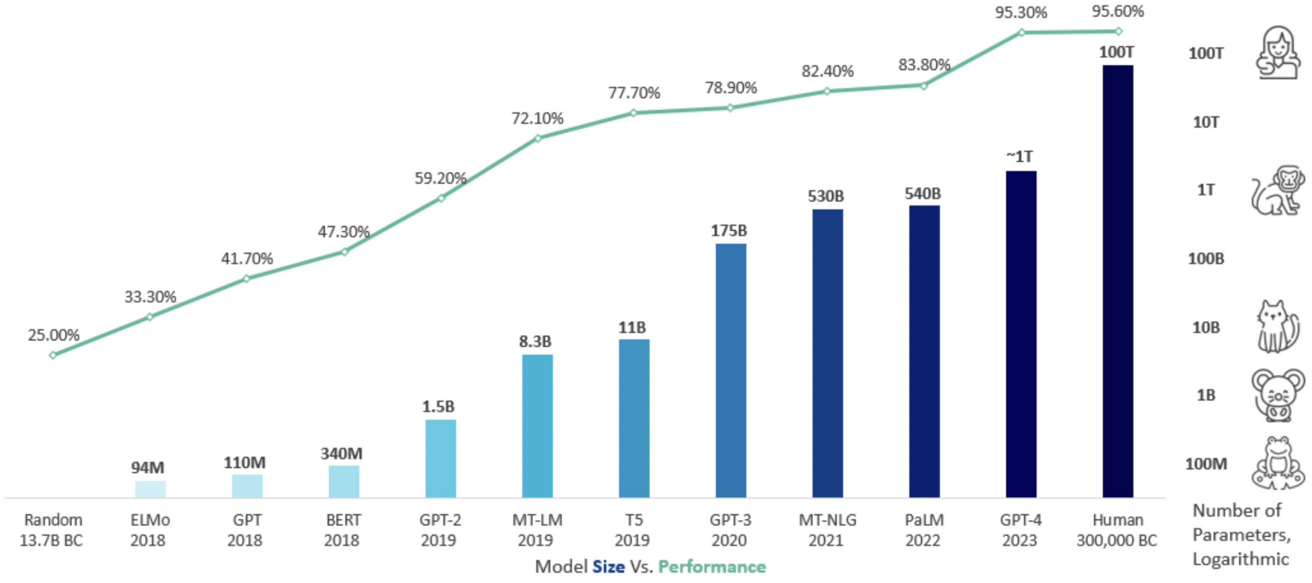
Innovation Ecosystems: Benchmarking AI Disruption

By Emmanuel Benhamou and Ash Tutika with Draup Intelligence

Artificial intelligence has become indispensable in the modern world, seamlessly integrating into countless industries. According to recent surveys, 50% of businesses now utilize AI in their operations, a dramatic rise from just 20% five years ago. This proliferation speaks to AI's immense value in optimizing processes, enabling data-driven decisions, and driving innovation.

From smart assistants in our pockets to self-driving cars on the streets, AI has progressed from science fiction to daily reality. Healthcare providers leverage AI for superior diagnostics and treatment planning.

Manufacturers implement AI on the factory floor to boost production. Across sectors, AI makes information more actionable and processes more efficient.



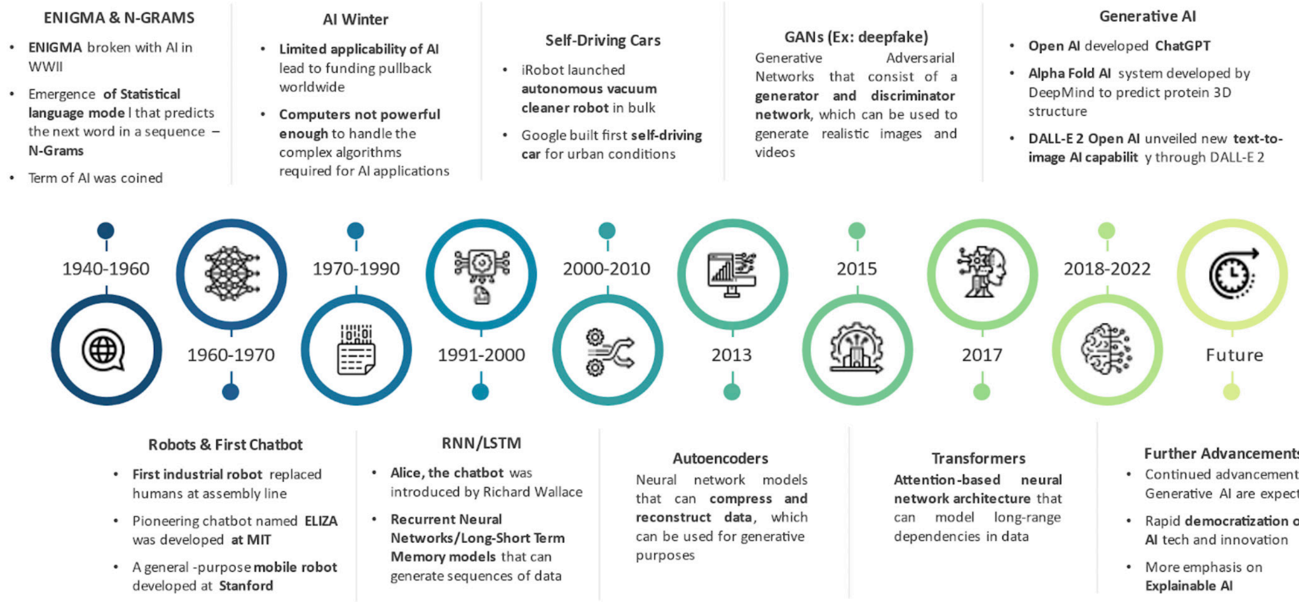
Source: Draup Intelligence, 2023, Model Size Vs. Performance

As AI capabilities continue to advance, we inch closer to a future where technology works symbiotically with humans, augmenting our capabilities in unprecedented ways. The transformational impact of AI on business and society is undeniable and ongoing. We have only begun to scratch the surface of AI's potential.

A profound transformation is unfolding, reshaping the global landscape across multiple dimensions. In this piece, we will look at the evolution of AI and the burgeoning new field of Generative AI, a critical subset enabling machines to emulate human-like outputs and create innovative content at the click of a button. We aim to provide a snapshot of the AI disruption, offering insights into the impact on industries, job roles, and various geographies, offering a clear perspective on the transformative changes and the emerging opportunities within the workforce. Further, a detailed analysis across multiple countries including Ireland, France, Singapore, Japan, and Canada, highlights the global footprint of AI, showcasing the unique challenges and advancements within each geographical context, and reinforcing the universal and far-reaching implications of AI and Generative AI in the contemporary world.

Evolution of Generative AI

Generative AI, a subset of AI, has emerged as a critical trend, allowing machines to create content and generate human-like outputs. Generative AI models, like GPT-4, contain a staggering 1 trillion parameters, enabling unprecedented capabilities.



ENIGMA & N-GRAMS

- **ENIGMA** broken with AI in WWII
- Emergence **of Statistical language mode** l that predicts the next word in a sequence – **N-Grams**
- Term of AI was coined

Robots & First Chatbot

- **First industrial robot** replaced humans at assembly line
- Pioneering chatbot named **ELIZA** was developed **at MIT**
- A general -purpose **mobile robot** developed at **Stanford**

AI Winter

- **Limited applicability of AI** lead to funding pullback worldwide
- **Computers not powerful enough** to handle the complex algorithms required for AI applications

RNN/LSTM

- **Alice, the chatbot** was introduced by Richard Wallace
- **Recurrent Neural Networks/Long-Short Term Memory models** that can generate sequences of data

Self-Driving Cars

- iRobot launched **autonomous vacuum cleaner robot** in bulk
- Google built first **self-driving car** for urban conditions

Autoencoders

Neural network models that can **compress and reconstruct data**, which can be used for generative purposes

GANs (Ex: deepfake)

Generative Adversarial Networks that consist of a **generator and discriminator network**, which can be used to generate realistic images and videos

Transformers

Attention-based neural network architecture that can model long-range dependencies in data

Generative AI

- **Open AI** developed **ChatGPT**
- **Alpha Fold AI** system developed by DeepMind to predict protein 3D structure
- **DALL-E 2 Open AI** unveiled new **text-to-image AI capabilit** y through DALL-E 2

Further Advancements

- Continued advancements in Generative AI are expected
- Rapid **democratization of AI** tech and innovation
- More emphasis on **Explainable AI**

Source: Draup Intelligence, 2023, Development of Generative AI

The evolution of Generative AI has seen remarkable progress, with key milestones shaping its development over the decades. It all began with the pioneering work of Alan Turing during World War II with the invention of the Enigma machine. While not strictly Generative AI, Turing's groundbreaking concept laid the foundation for understanding the potential of devices generating information.

Later, in the 1940s, the introduction of N-grams marked a significant advancement in language modeling. N-grams allowed computers to analyze sequences of words and predict the likelihood of certain word combinations, paving the way for early language generation experiments.

The following decades witnessed continuous growth in Generative AI capabilities. In the 1960s, researchers delved into rule-based systems for natural language processing (NLP). These early attempts, while limited, demonstrated the possibility of generating human-like responses to basic queries.

In the 1980s, researchers took on the challenge of semantic modeling, and created WordNet. Developed at Princeton University, Wordnet captures relationships between words such as synonyms, antonyms, hypernyms, and hyponyms for use in natural language processing.

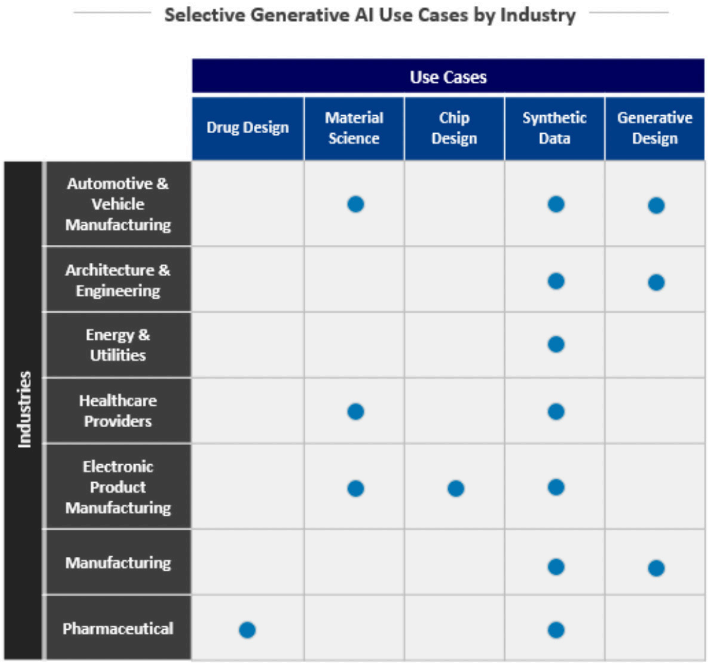
The 1990s saw the advent of probabilistic language models, such as Hidden Markov Models (HMMs) and n-gram-based statistical language models, which enabled more sophisticated language generation and improved context understanding.

Fast forward to the 21st century, Generative AI experienced a significant breakthrough with the introduction of transformer-based models. In 2018, the GPT (Generative Pre-trained Transformer) model pioneered large-scale language modeling and fine-tuning methods, revolutionizing natural language processing tasks. This led to the emergence of ChatGPT in 2021, a powerful conversational AI developed by OpenAI.

ChatGPT represents a culmination of advances in Generative AI, employing deep learning techniques and massive pre-training to generate human-like responses to a wide array of user queries. By leveraging large language models, such as GPT-4, ChatGPT exemplifies the current state-of-the-art in Generative AI and showcases the incredible potential of this technology for various applications across industries.

Impact on Industries

Generative AI has profoundly impacted several industries, driving transformative changes and enhancing operational efficiency. In Customer Support, deploying AI-driven chatbots and voice recognition technologies has revolutionized customer interactions, providing instant support and personalized assistance. Additionally, Predictive Analytics has empowered businesses to anticipate customer needs, enabling streamlined service delivery and improved customer satisfaction.



Selective Generative AI Use Cases by Industry

Industries	Drug Design	Material Science	Chip Design	Synthetic Data	Generative Design
Automotive & Vehicle Manufacturing		●		●	●
Architecture & Engineering				●	●
Energy & Utilities				●	
Healthcare Providers		●		●	
Electronic Product Manufacturing		●	●	●	
Manufacturing				●	●
Pharmaceutical	●			●	

Source: Gartner, Selective Generative AI Use Cases by Industry, 2023

In the Finance sector, Generative AI has significantly enhanced fraud detection and risk assessment processes. AI-powered algorithms analyze vast datasets in real-time, identifying suspicious activities and bolstering security measures. Moreover, AI-driven risk assessment models provide more accurate credit scoring and investment analysis, optimizing financial decision-making and portfolio management.

In Pharma R&D, Generative AI accelerates drug discovery and development. By generating potential compound structures and identifying promising candidates for clinical trials, AI-driven systems expedite research and reduce costs. This advancement could revolutionize the pharmaceutical industry and improve patient healthcare outcomes.

Generative AI's impact on various industries showcases the transformative potential of this technology. As it continues to evolve, we can expect further advancements and applications that will shape the future of industries worldwide.

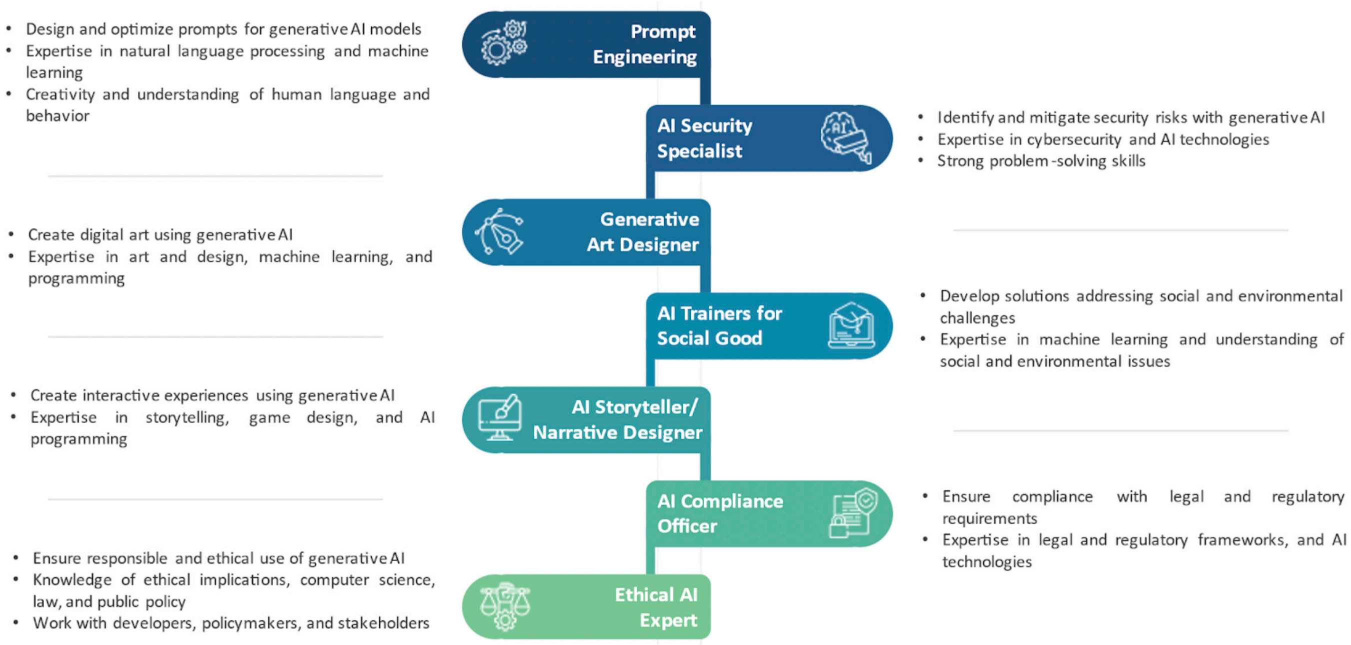
Generative AI has emerged as a critical driver of innovation and automation in robotics. AI-powered robots can now perform intricate tasks, previously limited to human capabilities, with incredible speed and accuracy. Generative AI algorithms enable these robots to learn from their experiences, adapt to new scenarios, and continuously improve their performance.

From manufacturing and logistics to healthcare and agriculture, AI-powered robots are revolutionizing industries by enhancing productivity, precision, and safety. As technology advances, we can expect to see even more sophisticated robotic systems collaborating seamlessly with humans, augmenting our capabilities, and unlocking new possibilities across various sectors.

Impact on Job Roles

The introduction of generative AI has led to several new roles, each with unique responsibilities and contributions to the rapidly evolving field of AI.

To start with, the role of a Prompt Engineer has become crucial. Prompt Engineers specialize in crafting effective prompts or instructions for AI models to generate desired outputs. Their expertise lies in understanding the nuances of language and context to elicit specific responses from AI systems. A well-designed prompt can lead to more accurate and contextually relevant outputs, making prompt engineers invaluable in tailoring AI applications to meet user needs effectively.



Prompt Engineering
- Design and optimize prompts for generative AI models
- Expertise in natural language processing and machine learning
- Creativity and understanding of human language and behavior

AI Security Specialist
- Identify and mitigate security risks with generative AI
- Expertise in cybersecurity and AI technologies
- Strong problem-solving skills

Generative Art Designer
- Create digital art using generative AI
- Expertise in art and design, machine learning, and programming

AI Trainers for Social Good
- Develop solutions addressing social and environmental challenges
- Expertise in machine learning and understanding of social and environmental issues

AI Storyteller/ Narrative Designer
- Create interactive experiences using generative AI
- Expertise in storytelling, game design, and AI programming

AI Compliance Officer
- Ensure compliance with legal and regulatory requirements
- Expertise in legal and regulatory frameworks, and AI technologies

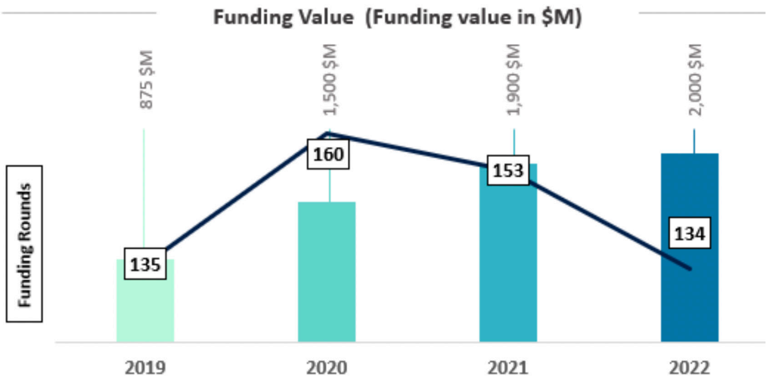
Ethical AI Expert
- Ensure responsible and ethical use of generative AI
- Knowledge of ethical implications, computer science, law, and public policy
- Work with developers, policymakers, and stakeholders

Source: Draup Intelligence, 2023, Generative AI impact on Job Roles

The need for AI Security Specialists will rise as AI systems become more sophisticated and pervasive. These experts focus on ensuring the security and integrity of generative AI models to prevent potential vulnerabilities and adversarial attacks. They work on developing robust defenses against negative inputs and unauthorized access to AI models, safeguarding sensitive data, and protecting AI systems from malicious exploitation.

Generative AI has also sparked the rise of Generative Art Designers. These professionals leverage AI algorithms to create unique and artistic outputs, pushing the boundaries of creativity in art and design. Generative Art Designers collaborate with AI models to produce captivating and innovative artworks, combining the capabilities of AI with human artistic expression.

Finally, the rise in adoption of AI has created the necessity and urgency for AI governance, compliance and risk professionals. New York City, for example, has recently passed a bill requiring regular audits of AI technologies used in human resources and hiring. The European Union also released the EU AI Act, with sweeping and groundbreaking laws for AI risk classification. With these regulatory changes



come an opportunity for AI risk and compliance professionals.

The emergence of new roles in generative AI is an exciting testament to the ongoing advancements in AI technologies and the collective efforts to shape AI's future in a manner that empowers and enriches humanity.

Countrywide Findings

Ireland

Dublin has emerged as a major artificial intelligence hub, though talent attraction faces challenges from salary gaps compared to other European cities. While Dublin boasts a vibrant technology scene, the city's median base salary for AI roles lags 22% behind London's compensation levels. This pay gap makes recruiting and retaining AI talent difficult despite Dublin's thriving ecosystem. However, demand for qualified AI professionals in the city remains robust, with a talent shortage of 42% - one of the highest talent gaps in Europe. To fully realize Dublin's potential as an AI leader, innovative approaches to talent development and competitive compensation strategies will be critical.

In response, the Irish government fosters an attractive environment for AI and data science talent, with initiatives like the AI Ireland strategy and establishing the National Centre for Applied Data Analytics and Machine Intelligence (CeADAR) in Dublin showcasing their commitment to building a robust AI ecosystem. The Digital Ireland Framework aims to achieve 75% AI integration in businesses by 2030.

Dublin has forged industry-academia partnerships. Renowned tech giants like Huawei, Intel, and IBM have established collaborations with prestigious institutions like Trinity College Dublin and Maynooth University. These partnerships facilitate knowledge exchange, research advancements, and skill development, driving innovation in the AI sector.

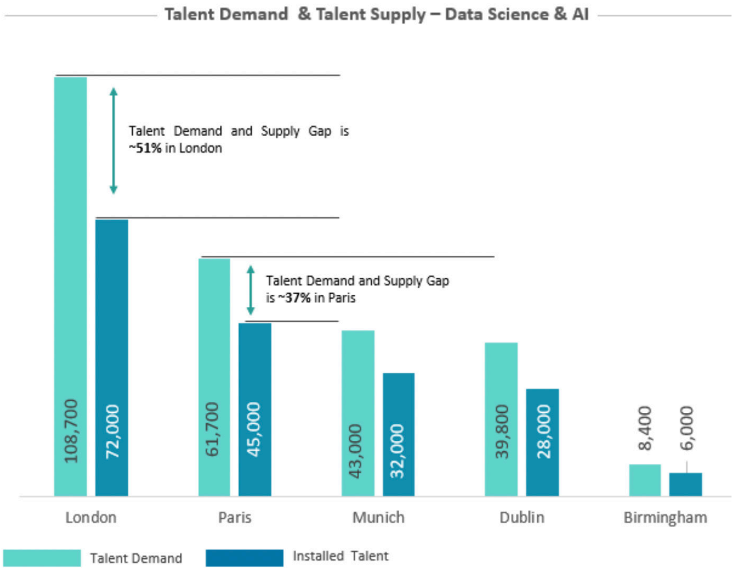
Dublin's thriving startup scene, hosting 9,433 start-ups, indicates its potential as an AI-driven entrepreneurship hotspot. Additionally, the presence of industry giants like Apple, Amazon, and IBM in nearby Cork creates a supportive ecosystem for data science and AI industries, attracting numerous smaller companies and startups to the region.

Ireland's innovation ecosystem has forged startups including Soapbox Labs, an AI speech recognition company for children, Boxever, a customer intelligence cloud for marketers, and Everseen, an AI anti-theft solution for retail. Venture capital funding into Irish startups rose by almost a third (32%) to a record €502 million in the first quarter of 2023, compared to €380 million in the same period last year, according to the Irish Venture Capital Association.

France

France's AI landscape is witnessing remarkable growth, with a surge in AI startups and research laboratories. As of 2023, there are 590 AI startups, representing a net increase of 24% compared to 2021. These startups have also seen a significant boost in funding, with the total amount raised doubling from €1.6 billion in 2021 to €3.2 billion in 2022.

The country boasts 81 AI research laboratories, the highest number among European nations, showcasing its commitment to advancing AI research and innovation. Interestingly, 15% of AI startups in France are solely dedicated to the health sector, highlighting the industry's potential to drive advancements in



healthcare and medical technologies.

French President, Emmanuel Macron, has taken proactive steps to foster the competitiveness of France's AI ecosystem. In June 2023, he announced a substantial investment of €500 million to further the growth and development of AI technologies in the country. This strategic move demonstrates the government's commitment to positioning France as a critical player in the global AI landscape.

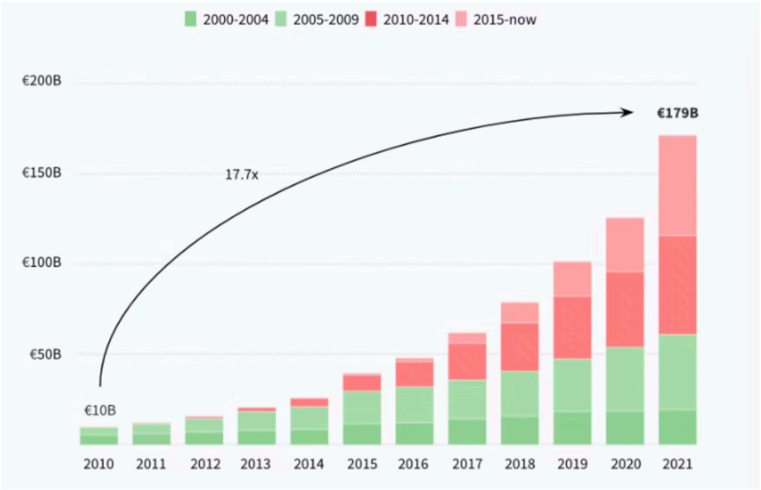
Over the past decade, France has emerged as a significant hotspot for venture capital (VC) investments within Europe. The French government's pro-innovation stance, combined with initiatives like La French Tech, has catalyzed an environment conducive to startups and attracted an increasing number from €1.6 billion in 2021 to €3.2 billion in 2022.

The country boasts 81 AI research laboratories, the highest number among European nations, showcasing its commitment to advancing AI research and innovation. Interestingly, 15% of AI startups in France are solely dedicated to the health sector, highlighting the industry's potential to drive advancements in healthcare and medical technologies.

French President, Emmanuel Macron, has taken proactive steps to foster the criticalcompetitiveness of France's AI ecosystem. In June 2023, he announced a substantial investment of €500 million to further the growth and development of AI technologies in the country. This strategic move demonstrates the government's commitment to positioning France as a critical player in the global AI landscape.

Over the past decade, France has emerged as a significant hotspot for venture capital (VC) investments within Europe. The French government's pro-innovation stance, combined with initiatives like La French Tech, has catalyzed an environment conducive to startups and attracted an increasing number of both domestic and international VC funds. Paris, in particular, has blossomed into a European tech hub.

Among the many French startups that have risen to prominence, "BlaBlaCar," a long-distance carpooling platform, stands out, having expanded its services across numerous countries. Another notable mention is "Doctolib," an online medical appointment scheduling service, which has become indispensable in many European healthcare systems. Additionally, "OVHcloud," a cloud computing company, has garnered international recognition, further validating the strength and potential of



Source: Dealroom, 2021, Combined Enterprise Value of French Startups the French startup ecosystem. These successes amplify France's position as a pivotal player in the European VC and startup arena.

Paris has also emerged as a major hub for artificial intelligence startups, with companies like Ledger, Contentsquare, Deezer, and Shift Technology driving innovation in the region. Ledger provides infrastructure for digital assets and has raised nearly $500 million from investors like Draper Esprit and Bpifrance. Contentsquare offers digital experience analytics, securing over $800 million from backers like Eurazeo and SoftBank. Music streaming service Deezer and AI-powered insurance technology provider Shift Technology likewise boast hundreds of millions in

funding and employees. These startups exemplify Paris' rising status as a center of AI development, leveraging the city's strong technology talent pool and investor networks. The success of these companies underscores the vast potential for continued AI growth and leadership from French startups.

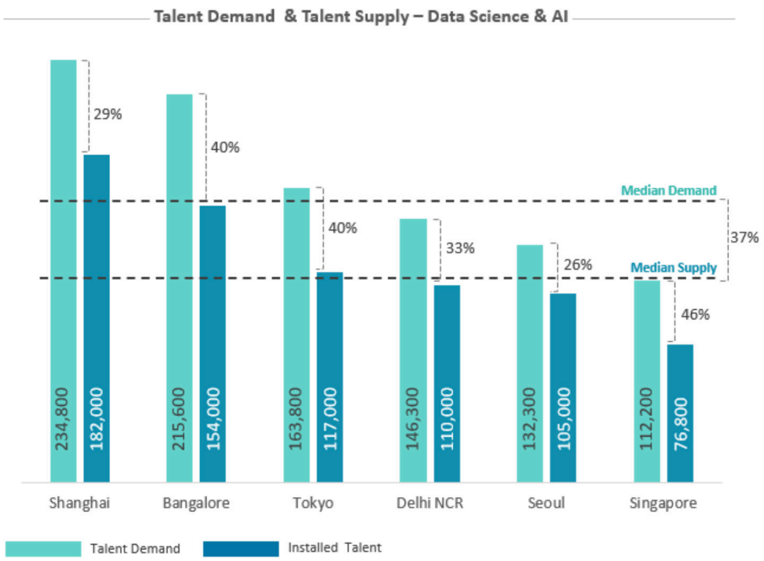
Combined Enterprise Value of French Startups

Moreover, France is also collaborating with Germany on the AI front. The joint funding of 5 AI projects worth 17.9 million euros signifies a robust effort to drive cross-border innovation and strengthen the ties between the two nations in artificial intelligence.

Despite the promising growth in the AI sector, France faces a talent demand-supply gap of 37%. The high demand for AI professionals presents challenges and opportunities for the country's workforce development and educational institutions.

Singapore

Singapore's AI landscape is experiencing rapid growth, but it also faces a significant talent demand-supply gap, the highest in Asia at 46%. To address this gap and attract top talent, Singapore has introduced initiatives like the Overseas Networks and Expertise Pass, which offers a five-year pass specifically designed for specific jobs in the tech industry. This pass aims to scout top talent across various sectors, emphasizing the tech industry and further bolstering Singapore's position as a regional tech hub.



AI is expected to play a pivotal role in Singapore's economic growth. By 2035, the manufacturing sector is projected to witness a substantial 40% increase in development with the integration of AI technologies. This promising growth trajectory also extends to the overall economy, as AI has the potential to boost Singapore's annual growth rate from 3.2% to an impressive 5.4% by 2035. Such growth translates to an additional USD 215 billion in gross value added, underscoring the transformative impact of AI adoption on Singapore's economic landscape.

Often dubbed the "Silicon Valley of Asia," Singapore has evolved into a leading startup hub in the region. The country's thriving ecosystem is valued at an impressive $21 billion, solidifying its position as the best startup ecosystem in South-East Asia and the 7th best startup ecosystem in the world.

In addition to its vibrant startup scene, Singapore offers competitive compensation for AI professionals, boasting the region's second-highest annual median base pay, trailing only behind Tokyo .In recent years, Singapore has positioned itself as a central hub for venture capital (VC) investments within the Asia-Pacific region. Fueled by supportive government policies, a robust financial ecosystem, and strategic geographical location, the city-state has witnessed an influx of both local and international VC funds. Several Singaporean startups have gained significant traction and success due to this increased investment. Among the standouts is "Grab," which began as a ride-hailing app and has since diversified into multiple domains, becoming Southeast Asia's first decacorn. "Carousell," a community marketplace for buying and selling items, is another success story that has expanded its presence across multiple countries in the region. Additionally, "Sea Group," which operates the popular gaming platform Garena and e-commerce site Shopee, has seen exponential growth and became the first Southeast Asian company to surpass a $100 billion market cap. These triumphs underscore Singapore's burgeoning reputation as a fertile ground for startups and the VC community.

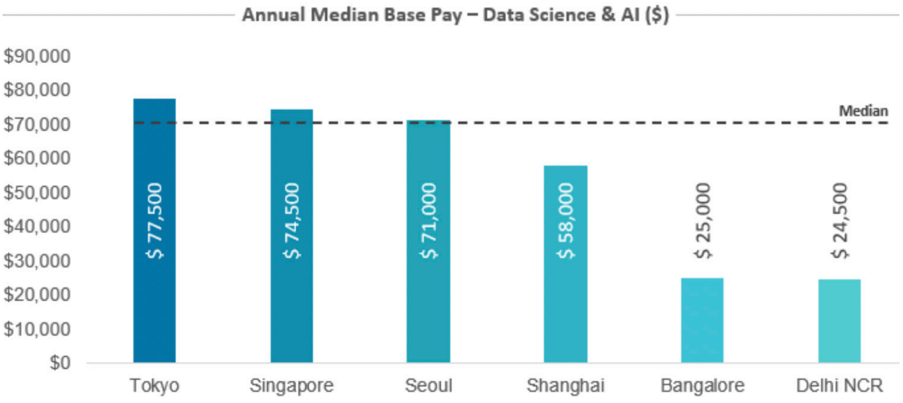
Japan

Japan's AI landscape is experiencing significant growth and innovation, but it also faces challenges in meeting the high demand for AI professionals. The country shares the second-highest supply-demand gap in Asia, with Bangalore standing at 40%.

Recognizing the potential talent shortage by 2030, the Japanese government is taking proactive measures to address this issue. They have targeted 250,000 AI experts from abroad through international education programs to attract and nurture talent.

Japan is at the forefront of embracing advanced technologies to transform its society. Through the "Society 5.0" initiative, the Japanese government envisions creating a "super-smart" society by integrating AI, IoT, and robotics into all aspects of daily life. The government invests heavily in research and development to support this ambitious vision.

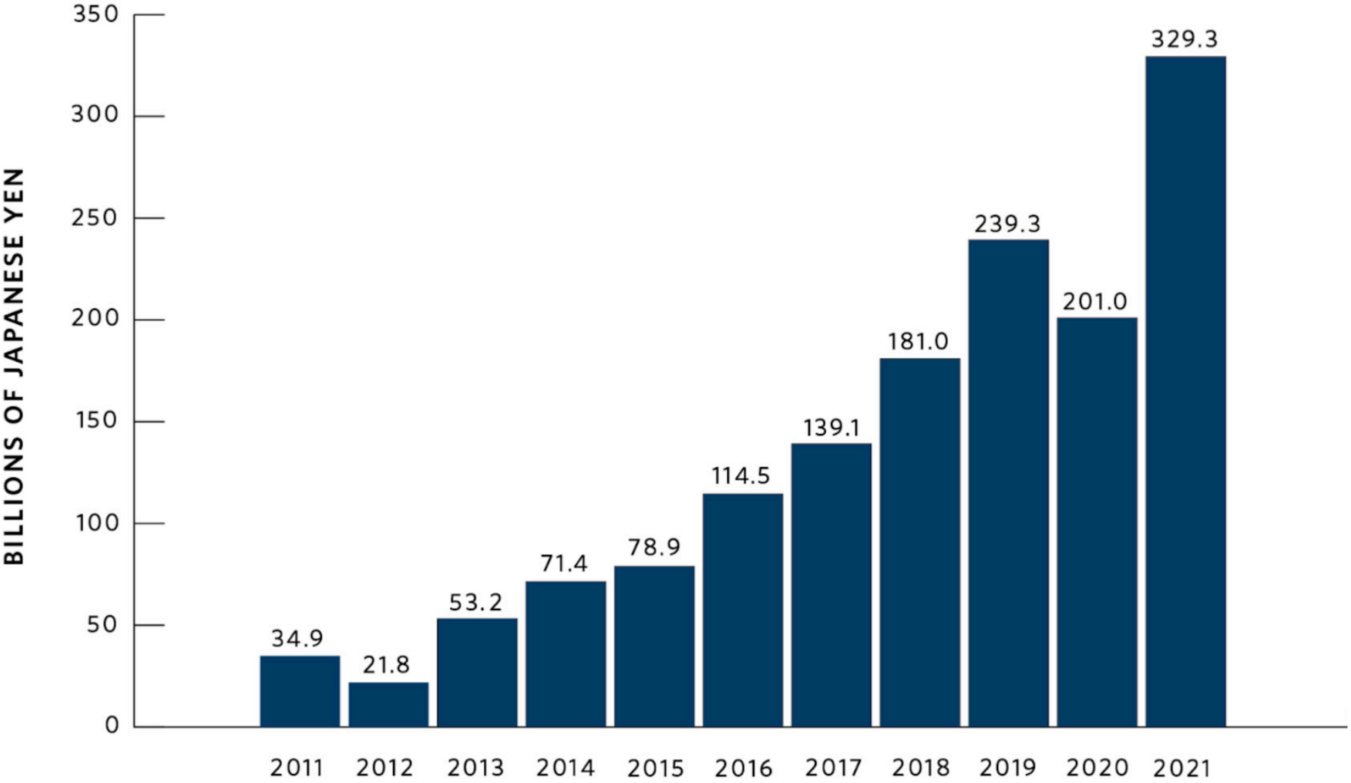
Japan's AI developments are not limited to the government sector but extend to fruitful collaborations between academia and industry. The University of Tokyo's Next Generation Artificial Intelligence Research Center has partnered with Toyota Central R&D Labs in the 'Intelligent Mobility Society Design, Social Cooperation Program.' This collaboration aims to drive advancements in AI applications related to intelligent mobility and pave the way for future transportation solutions.



Over recent years, Japan has witnessed a surge in venture capital (VC) investments, marking a significant departure from its traditionally conservative finance landscape. This growth can be attributed to a blend of government initiatives, a shift in corporate culture towards innovation, and increased interestfrom foreign investors recognizing the untapped potential of the Japanese startup ecosystem. Prominent startups that have thrived due to VC support include "Mercari," a consumer-to-consumer marketplace app

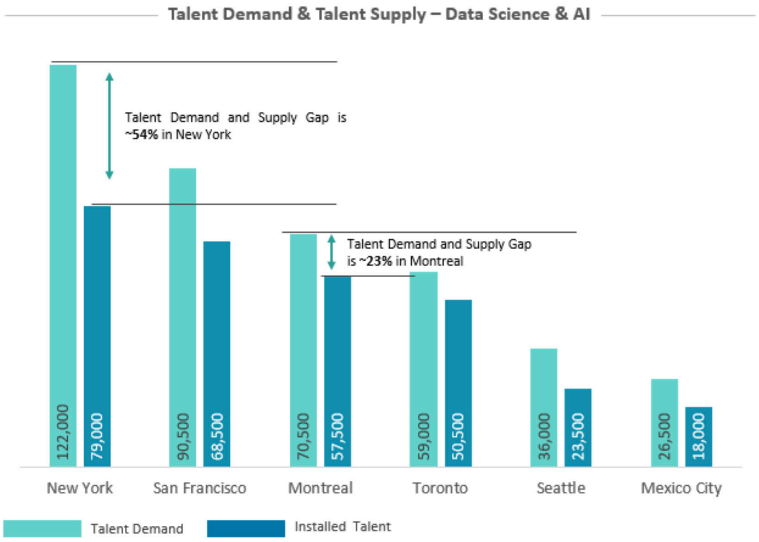
FIGURE 1

Japan's Venture Capital Investments



Source: "Japan Startup Funding 2021," Initial Enterprise, February 3, 2022, https://initial.inc/enterprise/resources/japanstartupfinance2021; "Japan Startup Funding 2020," Initial Enterprise

that became Japan's first unicorn, and "Raksul," an innovative printing and delivery service platform. Another noteworthy mention is "Preferred Networks," a deep learning startup that has garnered international attention for its advancements in artificial intelligence and its collaborations with major corporations. These successes not only illustrate the evolving dynamism of Japan's startup scene but also the increasing faith investors have in the country's entrepreneurial capabilities.Canada
Canada's AI sector has grown remarkably, with 600+ startups operating in the AI space by 2022. Toronto and Montreal are prominent AI hubs, generating 67%+ of these startups. Over 50% of funding goes to AI startups in Enterprise Infrastructure, HealthTech, and Enterprise Application industries, showing investor confidence in AI-driven innovations.

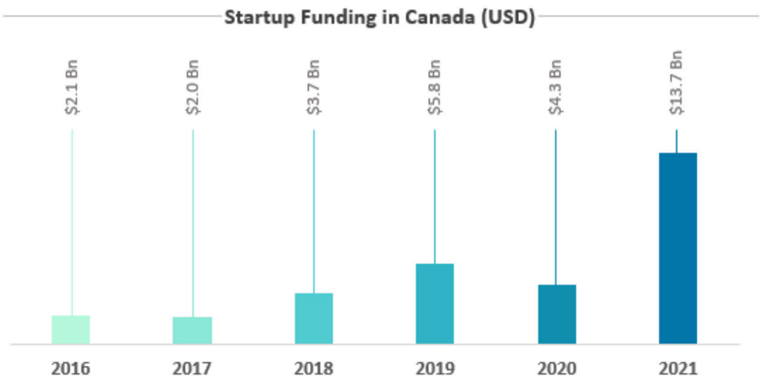


The Canadian AI research community is crucial in driving innovation and collaboration between academia and industry. MILA has established a team of applied research scientists dedicated to bridging the gap between research advances and practical applications.This approach fosters collaborations between MILA's ecosystem of researchers and industry partners, exemplified by the Orion-IBM project in partnership with IBM Canada. The project aims to accelerate the adoption of AI and machine learning using open-source technology, underscoring MILA's commitment to promoting real-world AI solutions through partnership and cooperation.

With approximately 290,000 tech workers, Toronto is only surpassed by Silicon Valley and New York City's tech workforce size. Moreover, Canada ranks 8th in skilled AI specialists globally, positioning the country as a critical player in the AI talent pool.

Situated in Quebec and stretching through the Montreal-Waterloo corridor, the Scale AI Cluster is revolutionizing the retail, manufacturing, transportation, infrastructure, and ICT sectors by seamlessly integrating artificial intelligence. This integration is fostering the creation of intelligent supply chains. Businesses are experiencing enhanced connectivity and efficiency, thanks to AI-driven tools that proactively predict product demands and optimize resource allocation in real time. As a result of this AI-driven synergy, the Scale AI Cluster is not only elevating Canada to the pinnacle of global exports but also ensuring that Canadian products and services with embedded AI capabilities pioneer in international markets.

Creative Destruction Labs, an accelerator program hosted out of the University of Toronto, Rotman School of Business, has incubated top AI companies in healthcare, enterprise software and conversational intelligence, including Ada and Darwin.ai.



The growth of AI in Canada has attracted significant investments from various companies, with more than 45 firms establishing AI research labs across the country. Notable companies like the Royal Bank of Canada, TD Bank, Microsoft, Nvidia, and Alphabet have invested in research and development centers in cities like Toronto, Montreal, Edmonton, and Vancouver.

Conclusion

The state of Artificial Intelligence is characterized by remarkable progress and widespread integration across various industries. AI has evolved from theoretical discussions to becoming an indispensable part of our daily lives, transforming businesses' operations and empowering innovation.

Generative AI, exemplified by models like GPT-4, has emerged as a critical trend, allowing machines to generate human-like content with unprecedented capabilities. The evolution of Generative AI, from Alan Turing's Enigma machine to transformer-based models, has shaped a future where AI augments human potential in remarkable ways.

Country-wide findings highlight the global impact of AI, with countries like Ireland, France, Singapore, Japan, and Canada witnessing significant growth and advancements in their AI ecosystems. As AI progresses, its responsible and strategic integration will pave the way for a future where technology and humanity coexist harmoniously, unleashing a new era of possibilities for the betterment of society.

KEY TAKEAWAYS

• AI adoption has proliferated rapidly, with 50% of businesses now utilizing AI in operations compared to just 20% five years ago. This proliferation speaks to AI's immense value in optimizing processes, enabling data-driven decisions, and driving innovation across sectors.

• Generative AI has emerged as a critical trend, allowing machines to generate human-like content with models like GPT-4 containing 1 trillion parameters. This represents the current state-of-the-art in natural language generation, exemplifying the incredible potential of AI.

• AI is profoundly impacting industries like customer support, finance, pharma R&D, and robotics by optimizing processes, bolstering security, expediting research, and enhancing productivity and precision. Across sectors, AI integration makes information more actionable and processes more efficient.

• New roles like Prompt Engineers, AI Security Specialists, Generative Art Designers, and AI compliance professionals have arisen due to advances in generative AI. Each role contributes unique expertise to shape the responsible development of AI.

• Dublin, France, Singapore, Japan, and Canada are experiencing surges in AI startups, research labs, investments, and government initiatives - establishing themselves as AI hubs. Strategic efforts by both public and private sectors are positioning these countries as global AI leaders.

• Countries face talent gaps in meeting AI professional demand, ranging from 37% in France to 46% in Singapore. To address shortages, governments are introducing policies to attract foreign talent and develop local skills.

• Collaborations between academia and industry are driving cross-sector innovation in AI across geographies. These partnerships enable knowledge sharing and practical applications of new AI capabilities.

• Advances in Generative AI exemplify the vast potential of AI to transform businesses and society. With responsible integration, AI promises an augmented human future where technology expands capabilities.

AI Will Change the Way Science Gets Done

Eric Schmidt

Biography

Eric Schmidt served as Chief Executive Officer of Google from 2001 to 2011, overseeing the company's rapid growth from startup to global technology leader. Schmidt provided strategic direction and management oversight during Google's expansion into search, email, mapping, video, mobile software and more. Revenues increased over 300x under his leadership.

Credited with bringing focus to Google's entrepreneurial culture, Schmidt guided the development of machine learning and artificial intelligence capabilities that became core competitive advantages. He helped establish Google's renown as an innovation pioneer that shaped the technology landscape.

Beyond Google, Schmidt has advocated for responsible development of AI, co-chairing the National Security Commission on AI and founding Schmidt Futures to bet early on people using technology for good. Schmidt's integral leadership powered Google's rise to tech superstardom, while his recent initiatives aim to ensure AI's benefits are shared broadly across society.

AI Will Chnge the Way Science Gets Done

By Eric Schmidt

It's yet another summer of extreme weather, with unprecedented heat waves, wildfires, and floods battering countries around the world. In response to the challenge of accurately predicting such extremes, semiconductor giant Nvidia is building an AI-powered "digital twin" for the entire planet.

This digital twin, called Earth-2, will use predictions from FourCastNet, an AI model that uses tens of terabytes of Earth system data and can predict the next two weeks of weather tens of thousands of times faster and more accurately than current forecasting methods.

Usual weather prediction systems have the capacity to generate around 50 predictions for the week ahead. FourCastNet can instead predict thousands of possibilities, accurately capturing the risk of rare but deadly disasters and thereby giving vulnerable populations valuable time to prepare and evacuate.

The hoped-for revolution in climate modeling is just the beginning. With the advent of AI, science is about to become much more exciting—and in some ways unrecognizable. The reverberations of this shift will be felt far outside the lab; they will affect us all.

Christoph Brugestedt/Science Photo Library

If we play our cards right, with sensible regulation and proper support for innovative uses of AI to address science's most pressing issues, AI can rewrite the scientific process. We can build a future where AI-powered tools will both save us from mindless and time-consuming labor and also lead us to creative inventions and discoveries, encouraging breakthroughs that would otherwise take decades.

AI in recent months has become almost synonymous with large language models, or LLMs, but in science there are a multitude of different model architectures that may have even bigger impacts. In the past decade, most progress in science has come through smaller, "classical" models focused on specific questions. These models have already brought about profound advances. More recently, larger deep-learning models that are beginning to incorporate cross-domain knowledge and generative AI have expanded what is possible.

Scientists at McMaster and MIT, for example, used an AI model to identify an antibiotic to combat a pathogen that the World Health Organization labeled one of the world's most dangerous antibiotic-resistant bacteria for hospital patients. A Google DeepMind model can control plasma in nuclear fusion reactions, bringing us closer to a clean-energy revolution. Within health care, the US Food and Drug Administration has already cleared 523 devices that use AI—75% of them for use in radiology.

Reimagining Science

Artificial intelligence is already transforming how some scientists conduct literature reviews. Tools like PaperQA and Elicit harness LLMs to scan databases of articles and produce succinct and accurate summaries of the existing literature—citations included.

Once the literature review is complete, scientists form a hypothesis to be tested. LLMs at their core work by predicting the next word in a sentence, building up to entire sentences and paragraphs. This technique makes LLMs uniquely suited to scaled problems intrinsic to science's hierarchical structure and could enable them to predict the next big discovery in physics or biology.

AI can also spread the search net for hypotheses wider and narrow the net more quickly. As a result, AI tools can help formulate stronger hypotheses, such as models that spit out more promising candidates for new drugs. We're already seeing simulations running multiple orders of magnitude faster than just a few years ago, allowing scientists to try more design options in simulation before carrying out real-world experiments.

Scientists at Caltech, for example, used an AI fluid simulation model to automatically design a better catheter that prevents bacteria from swimming upstream and causing infections. This kind of ability will fundamentally shift the incremental process of scientific discovery, allowing researchers to design for the optimal solution from the outset rather than progress through a long line of progressively better designs, as we saw in years of innovation on filaments in lightbulb design.

Moving on to the experimentation step, AI will be able to conduct experiments faster, cheaper, and at greater scale. For example, we can build AI-powered machines with hundreds of micropipettes running day and night to create samples at a rate no human could match. Instead of limiting themselves to just six experiments, scientists can use AI tools to run a thousand.

Scientists who are worried about their next grant, publication, or tenure process will no longer be bound to safe experiments with the highest odds of success; they will be free to pursue bolder and more interdisciplinary hypotheses. When evaluating new molecules, for example, researchers tend to stick to candidates similar in structure to those we already know, but AI models do not have to have the same biases and constraints.

Eventually, much of science will be conducted at "self-driving labs"—automated robotic platforms combined with artificial intelligence. Here, we can bring AI prowess from the digital realm into the physical world. Such self-driving labs are already emerging at companies like Emerald Cloud Lab and Artificial and even at Argonne National Laboratory.

Finally, at the stage of analysis and conclusion, self-driving labs will move beyond automation and, informed by experimental results they produced, use LLMs to interpret the results and recommend the next experiment to run. Then, as partners in the research process, the AI lab assistant could order supplies to replace those used in earlier experiments and set up and run the next recommended experiments overnight, with results ready to deliver in the morning—all while the experimenter is home sleeping.

Possibilities and Limitations

Young researchers might be shifting nervously in their seats at the prospect. Luckily, the new jobs that emerge from this revolution are likely to be more creative and less mindless than most current lab work.

AI tools can lower the barrier to entry for new scientists and open up opportunities to those traditionally excluded from the field. With LLMs able to assist in building code, STEM students will no longer have to master obscure coding languages, opening the doors of the ivory tower to new, nontraditional talent and making it easier for scientists to engage with fields beyond their own. Soon, specifically trained LLMs might move beyond offering first drafts of written work like grant proposals and might be developed to offer "peer" reviews of new papers alongside human reviewers.

AI tools have incredible potential, but we must recognize where the human touch is still important and avoid running before we can walk. For example, successfully melding AI and robotics through self-driving labs will not be easy. There is a lot of tacit knowledge that scientists learn in labs that is difficult to pass to AI-powered robotics. Similarly, we should be cognizant of the limitations—and even hallucinations—of current LLMs before we offload much of our paperwork, research, and analysis to them.

Companies like OpenAI and DeepMind are still leading the way in new breakthroughs, models, and research papers, but the current dominance of industry won't last forever. DeepMind has so far excelled by focusing on well-defined problems with clear objectives and metrics. One of its most famous successes came at the Critical Assessment of Structure Prediction, a biennial competition where research teams predict a protein's exact shape from the order of its amino acids.

From 2006 to 2016, the average score in the hardest category ranged from around 30 to 40 on CASP's scale of 1 to 100. Suddenly, in 2018, DeepMind's AlphaFold model scored a whopping 58. An updated version called AlphaFold2 scored 87 two years later, leaving its human competitors even further in the dust.

Thanks to open-source resources, we're beginning to see a pattern where industry hits certain benchmarks and then academia steps in to refine the model. After DeepMind's release of AlphaFold, Minkyung Baek and David Baker at the University of Washington released RoseTTAFold, which uses DeepMind's framework to predict the structures of protein complexes instead of only the single protein structures that AlphaFold could originally handle. More important, academics are more shielded from the competitive pressures of the market, so they can venture beyond the well-defined problems and measurable successes that attract DeepMind.

In addition to reaching new heights, AI can help verify what we already know by addressing science's replicability crisis. Around 70% of scientists report having been unable to reproduce another scientist's experiment—a disheartening figure. As AI lowers the cost and effort of running experiments, it will in some cases be easier to replicate results or conclude that they can't be replicated, contributing to a greater trust in science.

The key to replicability and trust is transparency. In an ideal world, everything in science would be open access, from articles without paywalls to open-source data, code, and models. Sadly, with the dangers that such models are able to unleash, it isn't always realistic to make all models open source.

In many cases, the risks of being completely transparent outweigh the benefits of trust and equity. Nevertheless, to the extent that we can be transparent with models—especially classical AI models with more limited uses—we should be.

The Importance of Regulation

With all these areas, it's essential to remember the inherent limitations and risks of artificial intelligence. AI is such a powerful tool because it allows humans to accomplish more with less: less time, less education, less equipment. But these capabilities make it a dangerous weapon in the wrong hands. Andrew White, a professor at the University of Rochester, was contracted by OpenAI to participate in a "red team" that could expose GPT-4's risks before it was released. Using the language model and giving it access to tools, White found it could propose dangerous compounds and even order them from a chemical supplier. To test the process, he had a (safe) test compound shipped to his house the next week. OpenAI says it used his findings to tweak GPT-4 before it was released.

Even humans with entirely good intentions can still prompt AIs to produce bad outcomes. We should worry less about creating the Terminator and, as computer scientist Stuart Russell has put it, more about becoming King Midas, who wished for everything he touched to turn to gold and thereby accidentally killed his daughter with a hug.

We have no mechanism to prompt an AI to change its goal, even when it reacts to its goal in a way we don't anticipate. One oft-cited hypothetical asks you to imagine telling an AI to produce as many paper clips as possible. Determined to accomplish its goal, the model hijacks the electrical grid and kills any human who tries to stop it as the paper clips keep piling up. The world is left in shambles. The AI pats itself on the back; it has done its job. (In a wink to this famous thought experiment, many OpenAI employees carry around branded paper clips.)

OpenAI has managed to implement an impressive array of safeguards, but these will only remain in place as long as GPT-4 is housed on OpenAI's servers. The day will likely soon come when someone manages to copy the model and house it on their own servers. Such frontier models need to be protected to prevent thieves from removing the AI safety guardrails so carefully added by their original developers.

To address both intentional and unintentional bad uses of AI, we need smart, well-informed regulation—on both tech giants and open-source models—that doesn't keep us from using AI in ways that can be beneficial to science. Although tech companies have made strides in AI safety, government regulators are currently woefully underprepared to enact proper laws and should take greater steps to educate themselves on the latest developments.

Beyond regulation, governments—along with philanthropy—can support scientific projects with a high social return but little financial return or academic incentive. Several areas are especially urgent, including climate change, biosecurity, and pandemic preparedness. It is in these areas where we most need the speed and scale that AI simulations and self-driving labs offer.

Government can also help develop large, high-quality data sets such as those on which AlphaFold relied—insofar as safety concerns allow. Open data sets are public goods: they benefit many researchers, but researchers have little incentive to create them themselves. Government and philanthropic organizations can work with universities and companies to pinpoint seminal challenges in science that would benefit from access to powerful databases.

Chemistry, for example, has one language that unites the field, which would seem to lend itself to easy analysis by AI models. But no one has properly aggregated data on molecular properties stored across dozens of databases, which keeps us from accessing insights into the field that would be within reach of AI models if we had a single source. Biology, meanwhile, lacks the known and calculable data that underlies physics or chemistry, with subfields like intrinsically disordered proteins that are still mysterious to us. It will therefore require a more concerted effort to understand—and even record—the data for an aggregated database.

The road ahead to broad AI adoption in the sciences is long, with a lot that we must get right, from building the right databases to implementing the right regulations, mitigating biases in AI algorithms to ensuring equal access to computing resources across borders.

Nevertheless, this is a profoundly optimistic moment. Previous paradigm shifts in science, like the emergence of the scientific process or big data, have been inwardly focused—making science more precise, accurate, and methodical. AI, meanwhile, is expansive, allowing us to combine information in novel ways and bring creativity and progress in the sciences to new heights.

KEY TAKEAWAYS

- **AI and the Evolution of Scientific Discovery**

Global Climate and AI's Potential Record-breaking heatwaves, wildfires, and floods have highlighted the need for improved weather predictions. Nvidia is working on an AI-powered digital twin of Earth named Earth-2, which uses the FourCastNet AI model. This model outpaces traditional systems by predicting weather with unprecedented speed and accuracy.

The Current State of AI in Science While large language models (LLMs) have recently made headlines, many scientific advancements still rely on smaller, focused models. Recent accomplishments include identifying antibiotics against dangerous bacteria, controlling nuclear fusion reactions, and significant applications in radiology.

- **Reimagining Science with AI**

The traditional scientific process—research, hypothesis, testing, data analysis, and conclusion—remains intact, but AI offers transformative approaches at every stage. LLMs are streamlining literature reviews, aiding in hypothesis formulation, and accelerating experimentation. The vision of "self-driving labs" suggests that much of the future scientific work will be automated, with AI lab assistants conducting research round-the-clock.

Benefits, Challenges, and Limitations of AI While AI offers increased efficiency, its broad adoption isn't without challenges. Potential pitfalls include the loss of intricate human knowledge that may not be easily passed on to AI. Additionally, AI's potential bias and the possibility of it being weaponized require serious consideration. The paper clip thought experiment illustrates the unintended consequences of an unchecked AI objective.

- **AI's Ecosystem: Open Source, Regulation, and Support**

With the growth of open-source models and resources, a balance between transparency and safety is crucial. Regulatory bodies must be well-informed and proactive, ensuring that AI tools are used responsibly without stifling innovation. Governments can bolster scientific research by supporting the creation of extensive, high-quality datasets and investing in areas of high societal benefit but low financial returns.

- **The Way Forward**

The integration of AI in scientific discovery presents an unprecedented opportunity to redefine research methodologies and outcomes. The journey ahead involves addressing challenges like database creation, regulation, bias mitigation, and resource allocation. But with careful planning and collaboration, AI can catalyze an expansive and transformative era in scientific discovery.

Despite Generative AI Buzz, Supervised Learning Will Create More Value Near Term

Written by Victor Dey of Venture Beat, featuring Andrew Ng

Biography

Andrew Ng is a leading figure in artificial intelligence and machine learning. He is the co-founder of multiple impactful organizations spawning from his pioneering work in AI. After earning bachelor's and master's degrees in computer science from Carnegie Mellon University, Ng obtained his PhD from UC Berkeley. He stayed on at Berkeley as an associate professor focused on machine learning and led the Google Brain project developing large scale artificial neural networks.

In 2011, Ng became a co-founder of Coursera, an online education platform providing universal access to world-class learning. Coursera has reached over 100 million learners worldwide. Ng was also a founding lead of the Google Brain deep learning project at Google. In 2011, he joined Baidu as chief scientist to build out Baidu's AI team into a world class research organization. In 2014, Ng left Baidu to launch deeplearning.ai, an AI startup focused on creating advanced education programs to teach deep learning techniques. Ng aimed to democratize access to AI knowledge beyond academics and technologists.

Ng is known for founding and leading the Google Brain project, developing the popular machine learning Coursera course, and his work growing AI teams at Google, Baidu and deeplearning.ai. His contributions have enabled broad dissemination of transformative AI concepts to students, developers and enterprises worldwide. Ng continues to advance the possibilities of AI through research papers, education initiatives, and new technologies lowering barriers to AI adoption. His pioneering work has helped fuel the era of deep learning across industries.

Despite Generative AI Buzz, Supervised Learning Will Create More Value Near Term

By Andrew Ng

One rarely gets to engage in a conversation with an individual like Andrew Ng, who has left an indelible impact as an educator, researcher, innovator and leader in the artificial intelligence and technology realms. Among the most prominent figures in AI, Andrew Ng is also the founder of DeepLearning.AI, co-chairman and cofounder of Coursera, and adjunct professor at Stanford University. In addition, he was chief scientist at Baidu and a founder of the Google Brain Project.

Venturebeat's encounter took place at a time in AI's evolution marked by both hope and controversy. Ng discussed the suddenly boiling generative AI war, the technology's future prospects, his perspective on how to efficiently train AI/ML models, and the optimal approach for implementing AI.

Momentum on the rise for both generative AI and supervised learning

VentureBeat: Over the past year, generative AI models like ChatGPT/GPT-3 and DALL-E 2 have made headlines for their image and text generation prowess. What do you think is the next step in the evolution of generative AI?

Andrew Ng: I believe generative AI is very similar to supervised learning, and a general-purpose technology. I remember 10 years ago with the rise of deep learning, people would instinctively say things like deep learning would transform a particular industry or business, and they were often right. But even then, a lot of the work was figuring out exactly which use case deep learning would be applicable to transform.

So, we're in a very early phase of figuring out the specific use cases where generative AI makes sense and will transform different businesses.

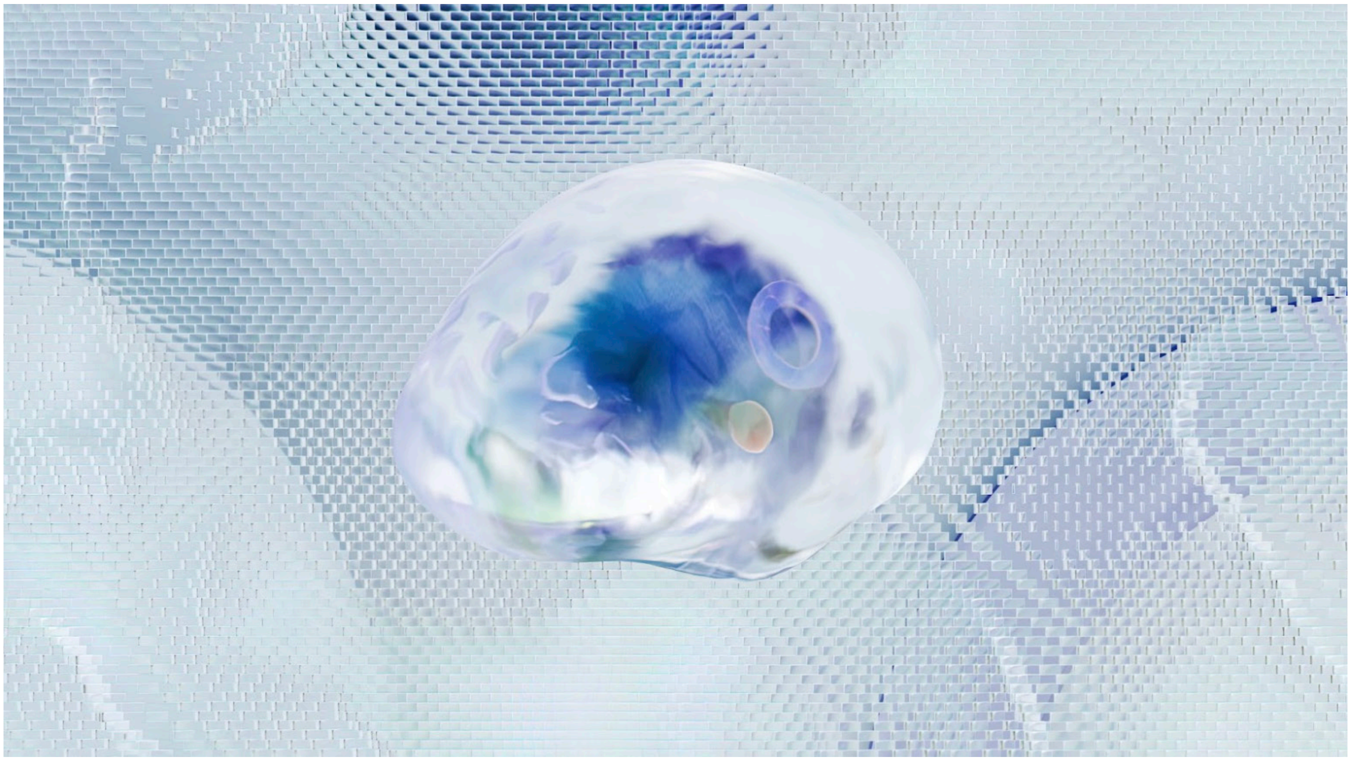
Also, even though there is currently a lot of buzz around generative AI, there's still tremendous momentum behind technologies such as supervised learning, especially since the correct labeling of data is so valuable. Such a rising momentum tells me that in the next couple of years, supervised learning will create more value than generative AI.

Due to generative AI's annual rate of growth, in a few years, it will become one more tool to be added to the portfolio of tools AI developers have, which is very exciting.

VB: How does Landing AI view opportunities represented by generative AI?

Ng: Landing AI is currently focused on helping our users build custom computer vision systems. We do have internal prototypes exploring use cases for generative AI, but nothing to announce yet. A lot of our tool announcements through Landing AI are focused on helping users inculcate supervised learning and to democratize access for the creation of supervised learning algorithms. We do have some ideas around generative AI, but nothing to announce yet.

Next-gen experimentation



Google Deepmind, 2023

VB: What are a few future and existing generative AI applications that excite you, if any? After images, videos and text, is there anything else that comes next for generative AI?

Ng: I wish I could make a very confident prediction, but I think the emergence of such technologies has caused a lot of individuals, businesses and also investors to pour a lot of resources into experimenting with next-gen technologies for different use cases. The sheer amount of experimentation is exciting, it means that very soon we will be seeing a lot of valuable use cases. But it's still a bit early to predict what the most valuable use cases will turn out to be.

I'm seeing a lot of startups implementing use cases around text, and either summarizing or answering questions around it. I see tons of content companies, including publishers, signed into experiments where they are trying to answer questions about their content.

Even investors are still figuring out the domain, so exploring further about the consolidation, and identifying where the roads are, will be an interesting process as the industry figures out where and what the most defensible businesses are.

I am surprised by how many startups are experimenting with this one thing. Not every startup will succeed, but the learnings and insights from lots of people figuring it out will be valuable.

VB: Ethical considerations have been at the forefront of generative AI conversations, given issues we're seeing in ChatGPT. Is there any standard set of guidelines for CEOs and CTOs to keep in mind as they start thinking about implementing such technology?

Ng: The generative AI industry is so young that many companies are still figuring out the best practices for implementing this technology in a responsible way. The ethical questions, and concerns about bias and generating problematic speech, really need to be taken very seriously. We should also be clear-eyed about the good and the innovation that this is creating, while simultaneously being clear-eyed about the possible harm.

The problematic conversations that Bing's AI has had are now being highly debated, and while there's no excuse for even a single problematic conversation, I'm really curious about what percentage of all conversations can actually go off the rails. So it's important to record statistics on the percentage of good and problematic responses we are observing, as it lets us better understand the actual status of the technology and where to take it from here.

Addressing roadblocks and concerns around AI

VB: One of the biggest concerns around AI is the possibility of it replacing human jobs. How can we ensure that we use AI ethically to complement human labor instead of replacing it?

Ng: It'd be a mistake to ignore or to not embrace emerging technologies. For example, in the near future artists that use AI will replace artists that don't use AI. The total market for artwork may even increase because of generative AI, lowering the costs of the creation of artwork.

But fairness is an important concern, which is much bigger than generative AI. Generative AI is automation on steroids, and if livelihoods are tremendously disrupted, even though the technology is creating revenue, business leaders as well as the government have an important role to play in regulating technologies.

VB: One of the biggest criticisms of AI/DL models is that they are often trained on massive datasets that may not represent the diversity of human experiences and perspectives. What steps can we take to ensure that our models are inclusive and representative, and how can we overcome the limitations of current training data?

Ng: The problem of biased data leading to biased algorithms is now being widely discussed and understood in the AI community. So every research paper you read now or the ones published earlier, it's clear that the different groups building these systems take representativeness and cleanliness data very seriously, and know that the models are far from perfect.

Machine learning engineers who work on the development of these next-gen systems have now become more aware of the problems and are putting tremendous effort into collecting more representative and less biased data. So we should keep on supporting this work and never rest until we eliminate these problems. I'm very encouraged by the progress that continues to be made even if the systems are far from perfect.

Even people are biased, so if we can manage to create an AI system that is much less biased than a typical person, even if we've not yet managed to limit all the bias, that system can do a lot of good in the world.

Getting real

VB: Are there any methods to ensure that we capture what's real while we are collecting data?

Ng: There isn't a silver bullet. Looking at the history of the efforts from multiple organizations to build these large language model systems, I observe that the techniques for cleaning up data have been complex and multifaceted. In fact, when I talk about data-centric AI, many people think that the technique only works for problems with small datasets. But such techniques are equally important for applications and works for problems with small datasets. But such techniques are equally important for applications and training of large language models or foundation models.

Over the years, we've been getting better at cleaning up problematic datasets, even though we're still far from perfect and it's not a time to rest on our laurels, but the progress is being made.

VB: As someone who has been heavily involved in developing AI and machine learning architectures, what advice would you give to a non-AI-centric company looking to incorporate AI? What should be the next steps to get started, both in understanding how to apply AI and where to start applying it? What are a few key considerations for developing a concrete AI roadmap?

Ng: My number one piece of advice is to start small. So rather than worrying about an AI roadmap, it's more important to jump in and try to get things working, because the learnings from building the first one or a handful of use cases will create a foundation for eventually creating an AI roadmap.

In fact, it was part of this realization that made us design Landing Lens, to make it easy for people to get started. Because if someone's thinking of building a computer vision application, maybe they aren't even sure how much budget to allocate. We encourage people to get started for free and try to get something to work and whether that initial attempt works well or not. Those learnings from trying to get into work will be very valuable and will give a foundation for deciding the next few steps for AI in the company.

I see many businesses take months to decide whether or not to make a modest investment in AI, and that's a mistake as well. So it's important to get started and figure it out by trying, rather than only thinking about [it], with actual data and observing whether it's working for you.

VB: Some experts argue that deep learning may be reaching its limits and that new approaches such as neuromorphic computing or quantum computing may be needed to continue advancing AI. What is your view on this issue?

Ng: I disagree. Deep learning is far from reaching its limits. I'm sure that it will reach its limits someday, but right now we're far from it.

The sheer amount of innovative development of use cases in deep learning is tremendous. I'm very confident that for the next few years, deep learning will continue its tremendous momentum. Not to say that other approaches won't also be valuable, but between deep learning and quantum computing, I expect much more progress in deep learning for the next handful of years.

KEY TAKEAWAYS

- **Generative AI:** The Next Frontier: Generative AI, with models like ChatGPT and DALL-E 2 at its forefront, is showing its prowess in image and text generation. However, Andrew Ng suggests we're just scratching the surface. Drawing a parallel with the rise of deep learning a decade ago, he emphasized the current challenge: pinpointing the exact niches where generative AI can be transformative. However, Ng also highlighted the enduring strength of supervised learning. Despite the buzz around generative AI, supervised learning, backed by accurately labeled data, will likely create more immediate value.

- **Landing AI's Vision:** Landing AI, founded by Ng, is a testament to his belief in the transformative power of AI. The company's primary focus is on custom computer vision systems. While they're navigating the generative AI terrain, their mainstay remains democratizing supervised learning, making algorithm creation accessible to a broader audience.

- **Uncharted Territories in Generative AI:** The rise of generative AI has catalyzed an era of experimentation. Ng's excitement stems from the sheer volume of exploration across various use cases, especially in content summarization and question-answering. As startups dive deep into this space, the industry is evolving to identify sustainable and lucrative business models.

- **Ethical Dimensions of AI:** Generative AI isn't devoid of ethical challenges. Ng expressed concerns about biases that can inadvertently creep into AI models, potentially leading to controversial outputs. He stresses the importance of keeping a close eye on the proportion of accurate versus problematic interactions. Furthermore, Ng touched upon another significant topic: the responsibility of business leaders and governments in ensuring AI doesn't drastically disrupt livelihoods. As AI continues to automate various functions, striking a balance between innovation and job preservation becomes paramount.

- **Emphasizing Diverse Training Data:** The foundation of any AI model is its training data. Ng emphasized the importance of unbiased, diverse datasets to ensure AI's successful application. Recognizing the flaws in existing models, he cited the AI community's ongoing efforts to rectify biases and gather inclusive data. The goal is an AI system with lesser bias than humans, which can significantly benefit society.

- **The Importance of Data Authenticity:** In a world where data is often referred to as the "new oil," ensuring its authenticity becomes vital. While there isn't a one-size-fits-all solution, Ng mentioned the strides made in refining data through data-centric AI. This approach is crucial not just for small datasets but also for training large foundational models.

- **AI in Traditional Businesses:** For businesses unfamiliar with the AI landscape, Ng's advice is clear: start small. Rather than getting bogged down by extensive planning, companies should take the plunge, gain hands-on experience, and learn from their endeavors. Such an approach can lead to more informed, strategic decisions in the longer run.

- **The Promise of Deep Learning:** As with all technologies, discussions around the potential limitations of deep learning have emerged. However, Ng's optimism about deep learning is palpable. He believes we're far from tapping its full potential. Even when pitted against promising technologies like quantum computing, Ng sees deep learning holding its ground for years to come.

Inside the Race to Build an Operating System for Generative AI

Written by Matt Marshall, Founder of Venture Beat, featuring Ashok Srivastava

Biography

Ashok Srivastava, Intuit's Chief Data Officer since 2019, is revolutionizing the company's approach to AI and data science. With over 25 years in the field, he has consistently harnessed data and AI to reshape businesses. Before joining Intuit, Ashok held pivotal roles at Verizon, focusing on big data analytics and machine learning. He also made significant contributions as Vice President of AI Research at AT&T Labs. At Intuit, Ashok's mission is to seamlessly integrate AI into both the company's products and its culture. He leads an elite team of data scientists and engineers, who have crafted advanced AI tools like virtual agents and recommendation engines, benefiting over 100 million users. A standout project under his guidance is the Generative AI Operating System (GenOS), a platform that orchestrates AI systems' interactions with enterprise resources, highlighting the transformative power of AI when paired with human ingenuity. Recognized for his unique blend of business savvy and profound AI knowledge, Ashok is a sought-after thought leader. He has penned over 50 research papers and is a Ph.D. alumnus from the University of Texas. Passionately curious, he's steering Intuit towards becoming a global, AI-centric platform.

Inside the Race to Build an Operating System for Generative AI

By Matt Marshall

Generative AI, the technology that can auto-generate anything from text, to images, to full application code, is reshaping the business world. It promises to unlock new sources of value and innovation, potentially adding $4.4 trillion to the global economy, according to a recent report by McKinsey.

But for many enterprises, the journey to harness generative AI is just beginning. They face daunting challenges in transforming their processes, systems and cultures to embrace this new paradigm. And they need to act fast, before their competitors gain an edge.

One of the biggest hurdles is how to orchestrate the complex interactions between generative AI applications and other enterprise assets. These applications, powered by large language models (LLMs), are capable not only of generating content and responses, but of making autonomous decisions that affect the entire organization. They need a new kind of infrastructure that can support their intelligence and autonomy.

Ashok Srivastava, chief data officer of Intuit, a company that has been using LLMs for years in the accounting and tax industries, told VentureBeat in an extensive interview that this infrastructure could be likened to an operating system for generative AI: “Think of a real operating system, like MacOS or Windows,” he said, referring to assistant, management and monitoring capabilities. Similarly, LLMs need a way to coordinate their actions and access the resources they need. “I think this is a revolutionary idea,” Srivastava said.

Credit: VentureBeat made with Midjourney

The operating-system analogy helps to illustrate the magnitude of the change that generative AI is bringing to enterprises. It is not just about adding a new layer of software tools and frameworks on top of existing systems. It is also about giving the system the authority and agency to run its own process, for example deciding which LLM to use in real time to answer a user's question, and when to hand off the conversation to a human expert. In other words, an AI managing an AI, according to Intuit's Srivastava. Finally, it's about allowing developers to leverage LLMs to rapidly build generative AI applications.

This is similar to the way operating systems revolutionized computing by abstracting away the low-level details and enabling users to perform complex tasks with ease. Enterprises need to do the same for generative AI app development. Microsoft CEO Satya Nadella recently compared this transition to the shift from steam engines to electric power. "You couldn't just put the electric motor where the steam engine was and leave everything else the same, you had to rewire the entire factory," he told Wired.

What does it take to build an operating system for generative AI?

According to Intuit's Srivastava, there are four main layers that enterprises need to consider.

First, there is the data layer, which ensures that the company has a unified and accessible data system. This includes having a knowledge base that contains all the relevant information about the company's domain, such as — for Intuit — tax code and accounting rules. It also includes having a data governance process that protects customer privacy and complies with regulations.

Second, there is the development layer, which provides a consistent and standardized way for employees to create and deploy generative AI applications. Intuit calls this GenStudio, a platform that offers templates, frameworks, models and libraries for LLM app development. It also includes tools for prompt design and testing of LLMs, as well as safeguards and governance rules to mitigate potential risks. The goal is to streamline and standardize the development process, and to enable faster and easier scaling.

Third, there is the runtime layer, which enables LLMs to learn and improve autonomously, to optimize their performance and cost, and to leverage enterprise data. This is the most exciting and innovative area, Srivastava said. Here new open frameworks like LangChain are leading the way. LangChain provides an interface where developers can pull in LLMs through APIs, and connect them with data sources and tools. It can chain multiple LLMs together, and specify when to use one model versus another.

Fourth, there is the user experience layer, which delivers value and satisfaction to the customers who interact with the generative AI applications. This includes designing user interfaces that are consistent, intuitive and engaging. It also includes monitoring user feedback and behavior, and adjusting the LLM outputs accordingly.

Intuit recently announced a platform that encompasses all these layers, called GenOS, making it one of the first companies to embrace a full-fledged gen OS for its business. The news got limited attention, partly because the platform is mostly internal to Intuit and not open to outside developers.

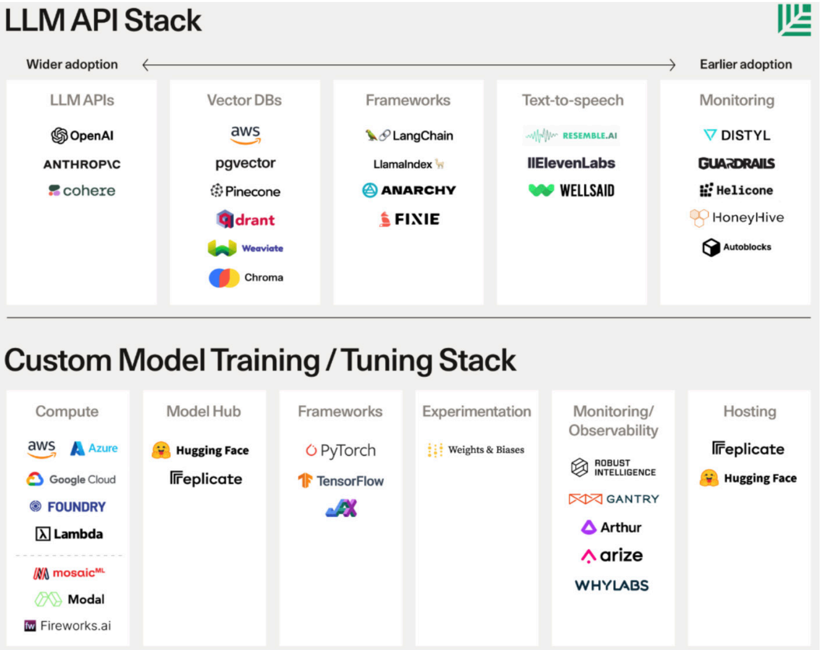
How are other companies competing in the generative AI space?

While enterprises like Intuit are building their own gen OS platforms internally, there is also a vibrant and dynamic ecosystem of open software frameworks and platforms that are advancing the state of the art of LLMs. These frameworks and platforms are enabling enterprise developers to create more intelligent and autonomous generative AI applications for various domains.

One key trend: Developers are piggy-backing on the hard work of a few companies that have built out so-called foundational LLMs. These developers are finding ways to affordably leverage and improve those foundational LLMs, which have already been trained on massive amounts of data and billions of parameters by other organizations, at significant expense. These models, such as OpenAI's GPT-4 or Google's PaLM 2, are called foundational LLMs because they provide a general-purpose foundation for generative AI. However, they also have some limitations and trade-offs, depending on the type and quality of data they are trained on, and the task they are designed for. For example, some models focus on text-to-text generation, while others focus on text-to-image generation. Some do better at summarization, while others are better at classification tasks.

Developers can access these foundational large language models through APIs and integrate them into their existing infrastructure. But they can also customize them for their specific needs and goals, by using techniques such as fine-tuning, domain adaptation and data augmentation. These techniques allow developers to optimize the LLMs' performance and accuracy for their target domain or task, by using additional data or parameters that are relevant to their context. For example, a developer who wants to create a generative AI application for accounting can fine-tune an LLM model with accounting data and rules, to make it more knowledgeable and reliable in that domain.

Another way that developers are enhancing the intelligence and autonomy of LLMs is by using frameworks that allow them to query both structured and unstructured data sources, depending on the user's input or context. For example, if a user asks for specific company accounting data for the month of June, the framework can direct the LLM to query an internal SQL database or API, and generate a response based on the data.



The New Language Model Stack, courtesy of Michelle Fradin and Lauren Reeder of Sequoia Capital

Unstructured data sources, such as text or images, require a different approach. Developers use embeddings, which are representations of the semantic relationships between data points, to convert unstructured data into formats that can be processed efficiently by LLMs. Embeddings are stored in vector databases, which are one of the hottest areas of investment right now. One company, Pinecone, has raised over $100 million in funding at a valuation of at least $750 million, thanks to its compatibility with data lakehouse technologies like Databricks.

Tim Tully, former CTO of data monitoring company Splunk, who is now an investor at Menlo Ventures, invested in Pinecone after seeing the enterprise surge toward the technology. "That's why you have 100 companies popping up trying to do vector embeddings," he told VentureBeat. "That's the way the world is headed," he said. Other companies in this space include Zilliz, Weaviate and Chroma.

What are the next steps toward enterprise LLM intelligence?

To be sure, the big-model leaders, like OpenAI and Google, are working on loading intelligence into their models from the get-go, so that enterprise developers can rely on their APIs, and avoid having to build proprietary LLMs themselves. Google's Bard chatbot, based on Google's PaLM LLM, has introduced something called implicit code execution, for example, that identifies prompts that indicate a user needs an answer to a complex math problem. Bard identifies this, and generates code to solve the problem using a calculator.

OpenAI, meanwhile, introduced function calling and plugins, which are similar in they can turn natural language into API calls or database queries, so that if a user asks a chatbot about stock performance, the bot can return accurate stock information from relevant databases needed to answer the question. Still, these models can only be so all-encompassing, and since they're closed they can't be fine-tuned for specific enterprise purposes. Enterprise companies like Intuit have the resources to fine-tune existing foundational models, or even build their own models, specialized around tasks where Intuit has a competitive edge — for example with its extensive accounting data or tax code knowledge base.

Intuit and other leading developers are now moving to new ground, experimenting with self-guided, automated LLM "agents" that are even smarter. These agents use what is called the context window within LLMs to remember where they are in fulfilling tasks, essentially using their own scratchpad and reflecting after each step. For example, if a user wants a plan to close the monthly accounting books by a certain date, the automated agent can list out the discrete tasks needed to do this, and then work through those individual tasks without asking for help. One popular open-source automated agent, AutoGPT, rocketed to more than 140,000 stars on Github. Intuit, meanwhile, has built its own agent, GenOrchestrator. It supports hundreds of plugins and meets Intuit's accuracy requirements.

The future of generative AI is here

The race to build an operating system for generative AI is not just a technical challenge, but a strategic one. Enterprises that can master this new paradigm will gain a significant advantage over their rivals, and will be able to deliver more value and innovation to their customers. They arguably will also be able to attract and retain the best talent, as developers will flock to work on the most cutting-edge and impactful generative AI applications.

Intuit is one of the pioneers and is now reaping the benefits of its foresight and vision, as it is able to create and deploy generative AI applications at scale and with speed. Last year, even before it brought some of these OS pieces together, Intuit says it saved a million hours in customer call time using LLMs.

Most other companies will be a lot slower, because they're only now putting the first layer — the data layer — in place. The challenge of putting the next layers in place will be at the center of VB Transform, a networking event on July 11 and 12 in San Francisco. The event focuses on the enterprise generative AI agenda, and presents a unique opportunity for enterprise tech executives to learn from each other and from the industry experts, innovators and leaders who are shaping the future of business and technology. Intuit's Srivastava has been invited to discuss the burgeoning GenOS and its trajectory. Other speakers and attendees include executives from McDonalds, Walmart, Citi, Mastercard, Hyatt, Kaiser Permanente, CapitalOne, Verizon and more. Representatives from large vendors will be present too, including Amazon's Matt Wood, VP of product, Google's Gerrit Kazmaier, VP and GM, data and analytics, and Naveen Rao, CEO of MosaicML, which helps enterprise companies build their own LLMs and just got acquired by Databricks for $1.3 billion. The conference will also showcase emerging companies and their products, with investors like Sequoia's Laura Reeder and Menlo's Tim Tully providing feedback.

I’m excited about the event because it’s one of the first independent conferences to focus on the enterprise case of generative AI. We look forward to the conversation.

KEY TAKEAWAYS

- Generative AI promises to unlock new innovation and value for enterprises, adding $4.4 trillion to the global economy. But companies face challenges in transforming processes and systems to embrace it.

- An "operating system" for generative AI is needed to orchestrate interactions between AI apps and enterprise assets. This OS needs capabilities for coordination, access to resources, autonomous optimization, etc.

- The OS has 4 main layers: data, development, runtime, and user experience. Companies like Intuit are building full stacks like this internally.

- Open frameworks are advancing the state of the art by improving foundational LLMs from OpenAI, Google etc. Techniques like fine-tuning and data augmentation customize models.

- New innovations like automated agents and context windows are making LLMs more intelligent and autonomous. Companies that master the new paradigms will gain a strategic advantage.

ML Ops in the Age of Generative AI

Krishna Gade

Biography

Krishna Gade is an experienced technology executive and serial entrepreneur. As CEO and founder of Fiddler, he is leading the development of a pioneering explainable AI system for monitoring machine learning models.

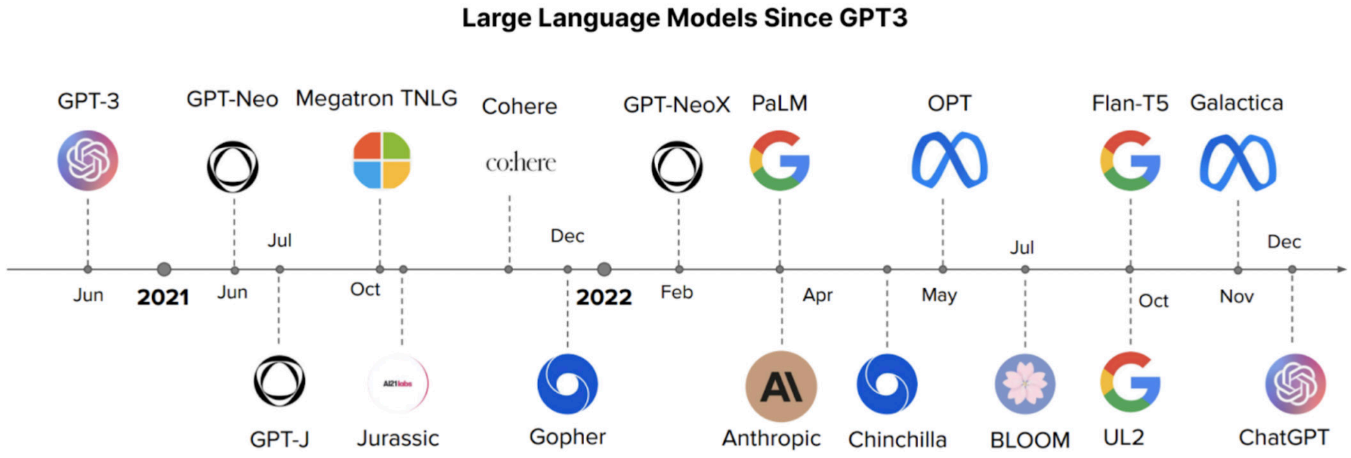
Previously, Krishna held senior engineering roles at major technology companies including Facebook, Twitter, Microsoft and Pinterest. At Microsoft, he was an early engineer on Bing's relevance algorithms and autosuggest features. He went on to lead data infrastructure teams at Twitter and Pinterest. Before launching Fiddler in 2018, Krishna served as an engineering manager for Facebook's news feed ranking team.

With over 15 years of experience across search, social media, infrastructure and AI, Krishna brings deep technical expertise in machine learning and data systems. As both an engineer and leader, he has driven innovation and built critical technologies at scale. Krishna holds multiple patents and frequently speaks at industry conferences. He co-founded Specialized Types, a startup advisory firm, and acts as an advisor to companies globally. Krishna studied computer science at MIT.

ML Ops in the Age of Generative AI

By Krishna Gade

The launch of GPT-3 and DALL-E ushered in the age of Generative AI and Large Language Models (LLM). With 175 billion parameters and trained on 45 TB of text data, GPT-3 was over 100x the 1.5 billion parameters of its predecessor. It validated OpenAI's hypothesis that models trained on larger corpora of data grew non-linearly in their capabilities. The next 18 months saw a cascade of innovation, with ever larger models, capped by the launch of ChatGPT at the tail end of 2022.



Source: Nazneen Rajani

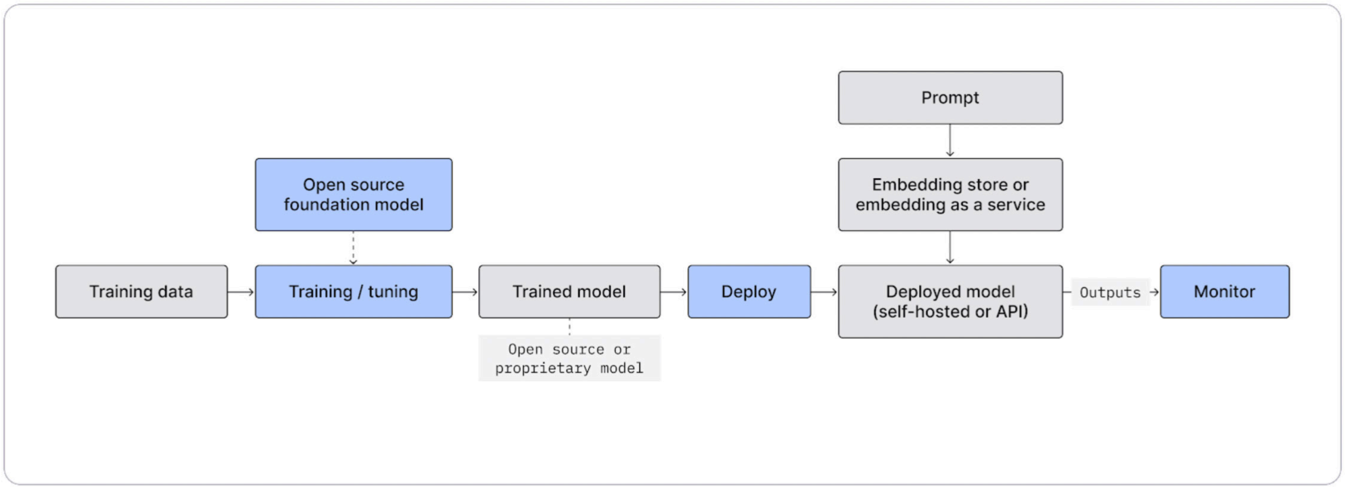
ChatGPT proved that AI is now poised to cross the technology chasm after decades of inching forward. All that remains is to operationalize this technology at scale. However, as we've seen with adoption of AI in general, the last mile is the hardest.

Path to Adopting Generative AI (LLMOps)

While Generative AI offers huge upside for enterprises, many blockers remain before it is used by a broad range of industries.

LLMs, especially the most recent models, have a large footprint and slow inference times, which require sophisticated and expensive infrastructure to run. Only companies with experienced ML teams with large resources can afford to bring models like these to market. OpenAI, Anthropic, and Cohere have raised billions in capital to productize these models.

Thankfully, the barrier to entry to productize Generative AI is quickly diminishing. Like ML Operations (MLOps), Generative AI needs an operationalized workflow to accelerate adoption. But which additional capabilities or tooling do we need to complete this workflow?


Prompt
Embedding store or embedding as a service
Open source foundation model
Training data
Training / tuning
Trained model
Deploy
Deployed model (self-hosted or API)
Outputs
Monitor
Open source or proprietary model

Generative AI Workflow (3rd party API, hosted proprietary or open source model)

Model Training

Recent AI breakthroughs are only possible by training with a large amount of advanced computational resources on a large corpora of data — prohibitively expensive for any company except ones with vast AI budgets. All LLMs from GPT-3 to the recently released LLaMa (Meta) have cost between $1M-$10M to train. For example, Meta's latest 65B LLaMa model training took 1,022,362 hours on 2048 NVidia A100-80GB's (approximately $4/hr on cloud platforms) costing approximately $4M. Besides the cost, building these model architectures demands an expert team of engineering and data science talent. For these reasons, new LLMs will be dominated by well capitalized companies in the near term.
Cost-efficient LLM training requires more efficient compute or new model architectures to unlock a sub-$10,000 cost for large models like the ones generating headlines today. This would accelerate a long tail of domain-specific use cases unlocking troves of data. With cloud providers dominating LLM training, one can hope these efficiencies develop over time.

Model Selection

Cost-effective model training is, however, not a deterrent to large scale Generative AI operationalization for two reasons (1) availability of open source that can be tuned (2) hosted proprietary models that can be invoked via API, i.e. AI-as-a-Service. For now, these are the two approaches that most AI teams will need to select from for their Generative AI use cases

1. Hosted Open Source Model - Majority of Generative AI innovation has come through models like Stable Diffusion which are open source. These "foundation models" will perform without needing any changes for the majority of use cases. However, they will still need to be finetuned with domain relevant data for use cases that require industry or function-specific context, i.e. medical chat, etc. We are seeing new fine tuning infrastructure being added at HuggingFace, Baseten Blueprint, etc. This tuning infrastructure is a key need for building foundational model "flavors".

2. Closed Source Model via API - While hosted open source models will be the norm in the long term given their lower cost and in-house ownership, OpenAI and Cohere have pioneered a new way to consume proprietary models via APIs. This approach will work well for a large number of AI teams that don't want to or don't have expertise to own these ML models. Eventually, companies similar to OpenAI will emerge. Instead of building new models, they will finetune foundational models for domain specific use cases and make them available to others via API.

Model Deployment

Model invocation cost is one of the biggest hurdles to adoption. The costs can be twofold: (1) inference speed and (2) expense driven by compute. For example, Stable Diffusion inference benchmarking shows a latency of well over 5 secs for 512 X 512 resolution images even on state of the art GPUs. Widespread adoption would require newer model architectures so that models can provide much faster inference speeds at lower deployment sizes while enabling comparable performance.

Coincidentally, companies are already making significant advances. Google AI recently introduced Muse, a new Text-To-Image approach that uses a masked generative transformer model instead of pixel-space diffusion or autoregressive models to create visuals. Not only does this run 10 times faster than Imagen and 3 times faster than Stable Diffusion, but it also accomplishes this with only 900 million parameters.

Embedding Ops

With Generative AI's focus on unstructured data, the representation of that data is a critical piece of the data flow. Embeddings represent this data and are typically the input currency of these models. How information is represented in these embeddings is a competitive advantage and can bring more efficient and effective inferences, especially for text models. In this sense, embeddings are equally (if not more) important than the models themselves.

Efficient embeddings are, however, non trivial to build and maintain. The rise of Generative AI APIs have also given rise to embedding APIs. Third party embedding APIs are bridging the gap in the interim by providing easy access to efficient embeddings at a cost. OpenAI, for example, provides an embeddings model, Ada, which costs $400 for every 1M calls for 1K tokens which can quickly add up at scale. In the long term, Generative AI deployments will need cheaper open source embedding models (eg. SentenceTransformers) that can easily be hosted to provide embeddings along with an embedding store, similar to a feature store, to manage them.

AI Monitoring and Safety

As we've discussed, Generative AI is not cheap. On OpenAI's Foundry platform, running a lightweight version of GPT-3.5 will cost $78,000 for a three-month commitment or $264,000 over a one-year commitment. To put that into perspective, one of Nvidia's recent-gen supercomputers, the DGX Station, runs $149,000 per unit. Therefore, a high performance and low cost Generative AI application will need comprehensive monitoring infrastructure irrespective of whether the models are self-hosted or are being invoked via API from a third party.

It's well known that model performance degrades over time, known as model drift, resulting in models losing their predictive power, failing silently, or harboring risks for businesses and their customers. Companies typically employ model monitoring to ensure their ML powered businesses are not impacted by the underlying model's operational issues. Like other ML models, Generative AI models can bring similar and even new risks to users.

The most common problem plaguing these models is correctness of the output. Some prominent examples have been both Google Bard and Microsoft Bing's errors and AI's flawed generation of human fingers. The impact of inaccuracies is amplified for critical use cases that could lead to potential harm eg. incorrect or misleading medical information, encouraging self-harm etc. These incorrect outputs need to be recorded to improve the model's quality.

Prompts are the most common way end users interact with Generative AI models, and the second biggest issue is prompt iteration to reach a desired output. Some prompts might give ineffective outputs while other prompts might not have sufficient data to generate a good output. In both cases, this results in customer dissatisfaction that needs to be captured to assess if the model is performing poorly in some areas after its release.

Generative AI models can also encounter several other operational issues. Data or embeddings going into the models can shift over time impacting model performance — this is typically evaluated with comparison metrics like data drift. Model bias and output transparency are lingering concerns for all ML models and are especially exacerbated with large data and complex Generative AI models. Performance might change between versions, so customers need to run tests to find the most effective models. Costs can catch up quickly, so monitoring expenses of these API calls and finding the most effective provider is important. Safety is another new concern either from the model's objectionable outputs or from the user's adversarial inputs. Monitoring solutions can provide Generative AI users visibility into all these operational challenges.

The onset of Generative AI will see an explosion of API driven users given the ease of API integrations, soon followed by a rapid increase of hosted custom Generative AI models. Infrastructure tooling will therefore follow a similar arc that will enable the "AI-as-a-service" use case first and the hosted custom AI use case next. Over time the maturation of this infrastructure in training, tuning, deploying, and monitoring will bring Generative AI to the wider masses.

KEY TAKEAWAYS

- **Path to Adopting Generative AI:**
  - Generative AI like GPT-3 and DALL-E offer huge potential, but barriers remain for broad enterprise adoption. The large footprint and slow inference of LLMs require expensive infrastructure and skilled teams. An MLOps-like workflow is needed to operationalize and accelerate adoption.

- **Generative AI Workflow:**
  - Model training requires massive compute resources and data, costing millions for companies like Meta and OpenAI. Two approaches are using hosted open source models or closed source models via API. Models still need domain-specific fine-tuning. Companies will offer hosted tuned models as a service.

- **Model Deployment:**
  - Inference speed and expense are hurdles to adoption. Advances are being made in model architectures for faster and lower-cost deployment, like Google's Muse. Efficient embeddings are key for unstructured data input. New embedding APIs are bridging gaps.

- **AI Monitoring:**
  - Monitoring is critical for high-cost API usage and model performance issues like incorrect outputs and ineffective prompts. Other issues include data drift, bias, and performance across versions. Solutions provide visibility into operational challenges.

- **Overall:**
  - An operationalized workflow for training, deployment, and monitoring is key. As infrastructure matures, it will bring Generative AI into wider use. Transition expected from API use to custom hosted models. Infrastructure advances will enable this AI-as-a-service model.

Enhancing Data to Boost Machine Learning Model Performance

Avi Weiss

Biography

Avi Weiss is the founder and CEO of Datomize, a company he started in 2020. As CEO, Weiss leads Datomize in its mission to reinvent statements and bills for enterprise-level companies. Before founding Datomize, Weiss gained over 20 years of industry experience. He served as CEO of ActivePath from 2015-2020, helping the company provide interactive statements and bills to improve customer engagement. Prior to ActivePath, Weiss co-founded ObserveIT in 2007 and served as COO for over 12 years. His previous leadership experience includes serving as CEO of Cloverleaf Communications and general manager of Tradeum/VerticalNet's Israel operations.

With over 35 years of technology industry experience, Weiss is an expert in customer communications management. Since 1983, he has taken on various roles from project manager to founder and CEO. His technical knowledge combined with business acumen has enabled Weiss to recognize market opportunities and successfully build innovative companies like Datomize and ActivePath. Under his leadership, Datomize aims to transform how enterprises connect with their customers through statements, bills and other communications.

Enhancing Data to Boost Machine Learning Model Performance

Sigal Shaked

Biography

Sigal Shaked is the Co-Founder and CTO of Datomize, an AI-powered data generation platform founded in 2019. As CTO, Sigal oversees the company's technology strategy and data science research.Prior to co-founding Datomize, Sigal served as Head of Data Science at a hedge fund from 2018-2019. In this role, she led data science solutions across various projects, conducted applied research, and managed the data science team.

From 2007-2018, Sigal was a researcher at Deutsche Telekom, where she worked on projects involving network traffic generation, call detail record fabrication, and location prediction. She has also been a lecturer at Ben-Gurion University since 2006, teaching courses in data science, databases, and more. Sigal holds a Ph.D. in Software and Information Systems Engineering from Ben-Gurion University, where she researched sequence preserving data generation techniques for privacy preservation using machine learning and deep learning. She also earned an M.Sc. in Information Systems Engineering from Ben-Gurion University. Through her research and work at Datomize, Sigal is an expert in leveraging AI and machine learning for responsible and privacy-preserving data generation. She is passionate about advancing data science capabilities while protecting data privacy.

Enhancing Data to Boost Machine Learning Model Performance

By Avi Weiss and Sigal Shaked

The Primacy of Data-Centric Approaches in the Age of Generative AI

For the past two decades, the prevailing approach to artificial intelligence (AI) has been model-centric. This strategy focused on crafting machine learning models adept at bridging data gaps. The goal was to iteratively refine these models to peak performance, all while leaving the data untouched.

However, as we advanced into 2021, a new perspective emerged: the data-centric approach. This methodology underscores the importance of optimizing training data. Instead of constantly tweaking the model, the emphasis is on elevating the quality of the data, keeping the model or code consistent.

Last year marked a significant milestone with the rise of generative AI. This technology garnered attention for its prowess in generating diverse content - from images and videos to emails and program codes. These tools offered novel content, translations, sentiment analyses, and more.

Yet, the real innovation is unfolding behind the scenes, in the realm of structured data. Here, generative models are making strides in data augmentation, balancing, imputation, and cleansing. With its proven capabilities, it's compelling to harness generative AI to enhance data quality. After all, generative AI offers the potential to mold the data we gather into the precise information we desire.

Before delving into the enhancements of machine learning models, it's pivotal to understand the intricate relationship between a machine learning task and its corresponding data. This connection is direct and crucial. Data, often riddled with imbalances, noise, or biases, can lead to predictions that are not only inaccurate but potentially detrimental. Models tainted with biases, be it racial, gender, or geographic, can pose significant risks to organizations.

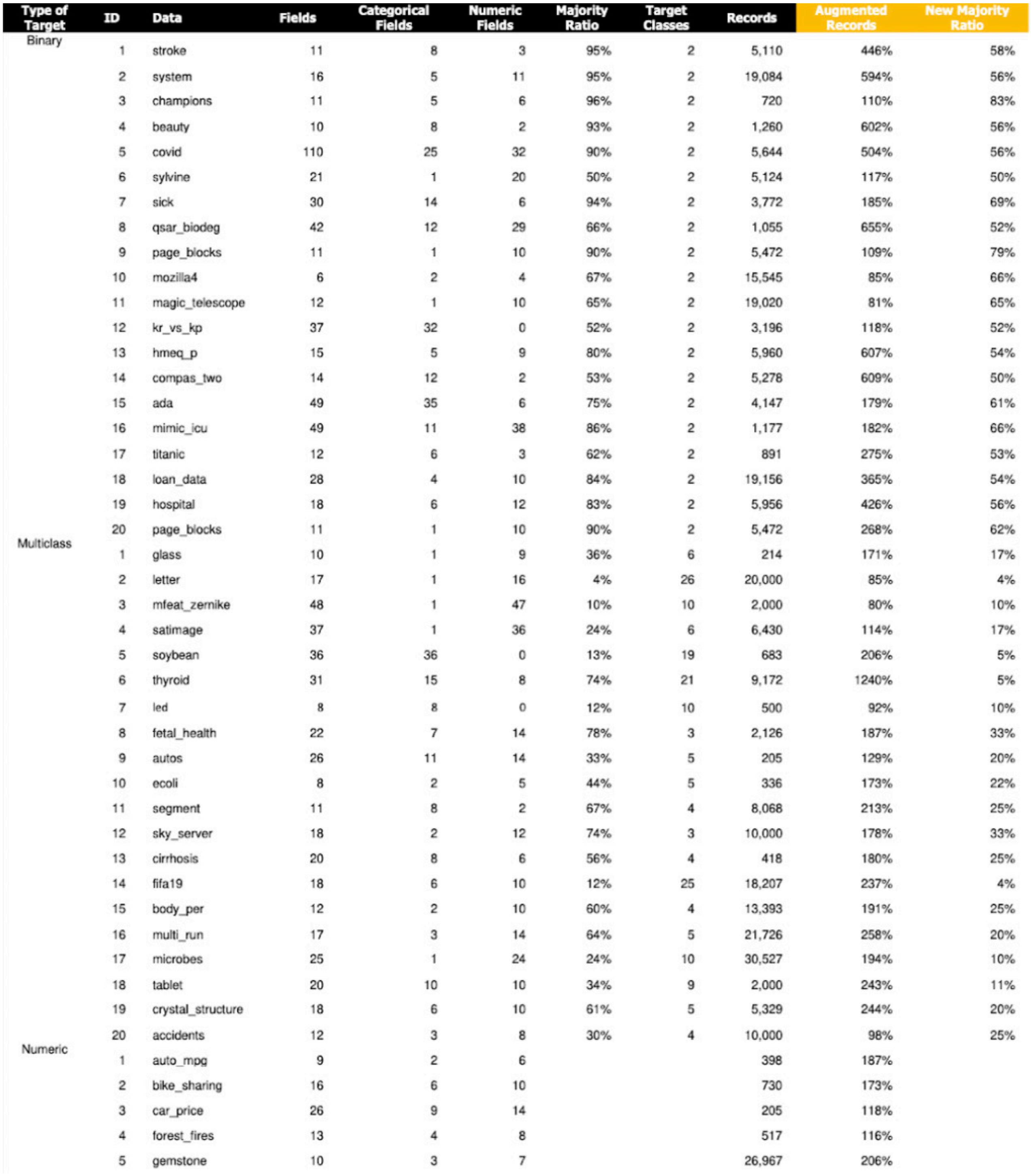
The solution? Optimize the data from the outset. By addressing these flaws before the training phase, we pave the way for machine learning predictions that are more accurate, efficient, and reliable.

Machine learning encompasses various tasks. In classification, models categorize data, whereas in regression, they predict specific values. While regression tasks offer a broad spectrum of potential outcomes, classification, especially binary classification, often grapples with imbalances. Tasks like fraud detection or disease prediction are particularly challenging due to the underrepresentation of certain outcomes. Techniques like undersampling and oversampling are pivotal in addressing these imbalances. Furthermore, numeric prediction tasks, which yield continuous outcomes, also present their unique challenges. Not all potential outcomes might be represented in the data, necessitating optimization techniques to enrich the data.

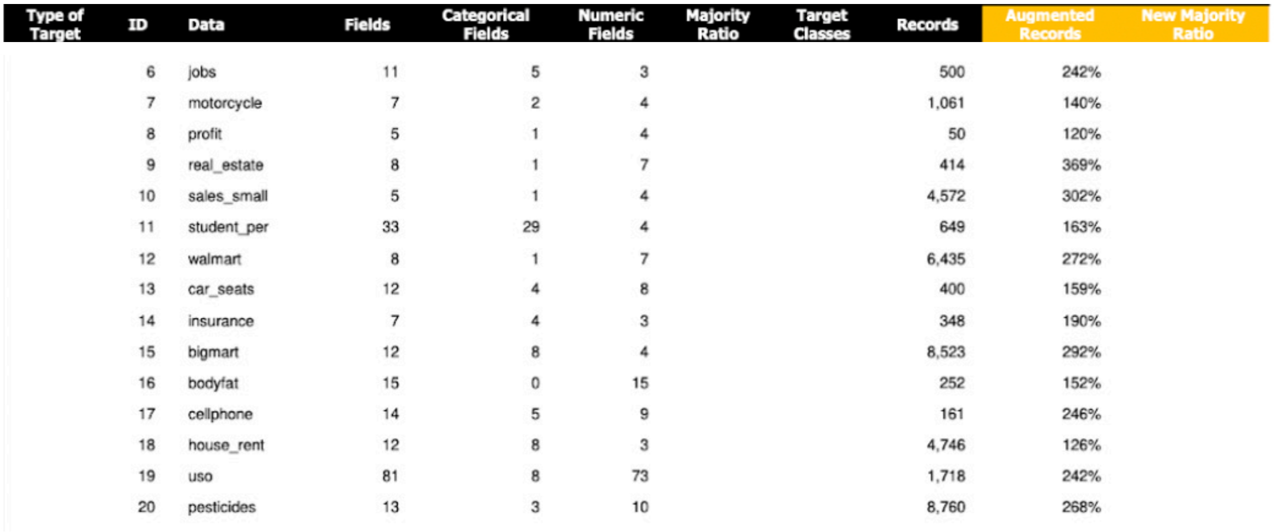
This document underscores the potency of our data-centric approach, bolstered by generative AI. By prioritizing the optimization of training data, we unveil the potential for more precise and accurate prediction models.

Our Data Enhancement Process

In our data enhancement process, a variational autoencoder (VAE) is used to transform the input data into a lower-dimensional feature space, where new representations are generated using techniques like crossover and SMOTE. These are later decoded back to the original feature space and serve as augmented records for the source data. Grid search is performed on a validation set to optimize various process-related parameters like the augmentation and balancing ratios, given a specific evaluation metric. Experiment We evaluated our data enhancement process on 60 small- to medium-sized datasets (datasets with hundreds to tens of thousands of records), of which 20 were related to a binary target classification task, 20 to a multiclass target classification task, and 20 to a numeric target regression task. The datasets and their metadata are presented in Table 1.



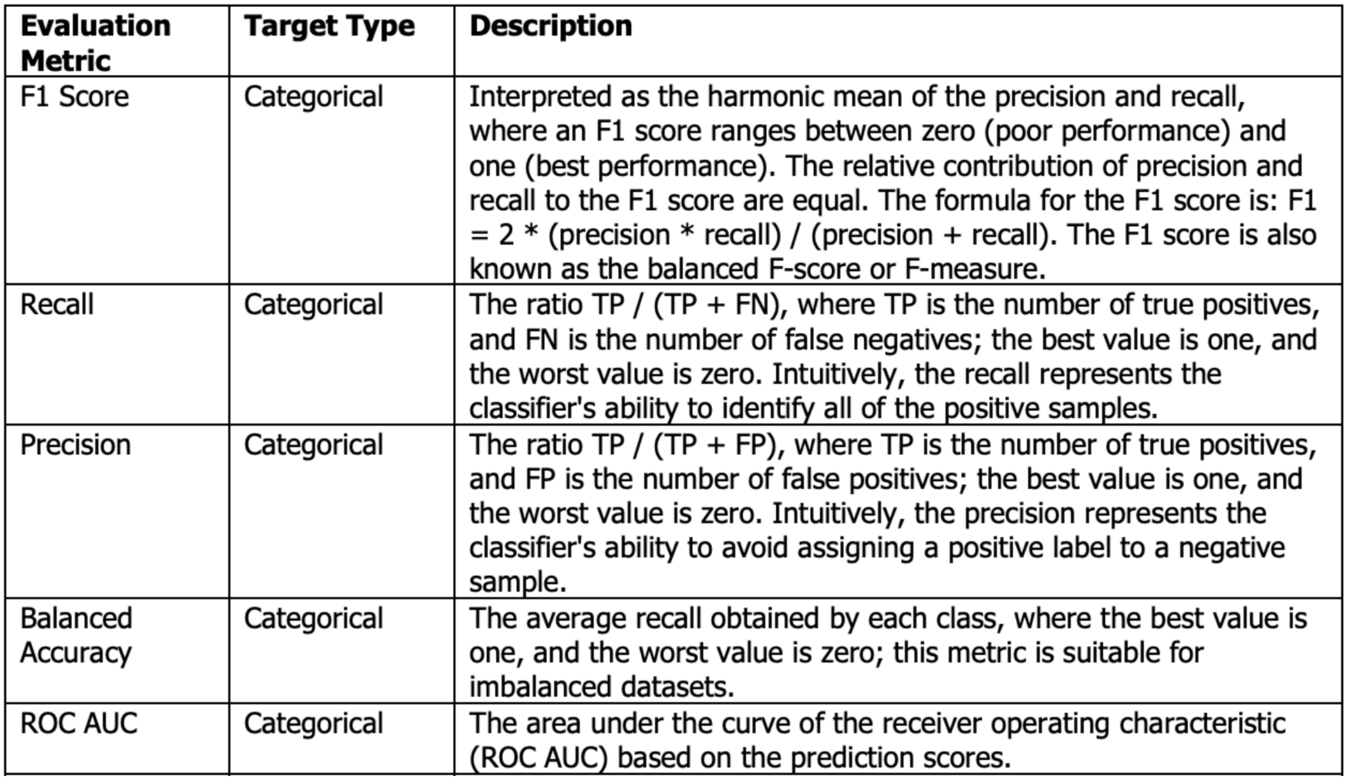
Type of Target	ID	Data	Fields	Categorical Fields	Numeric Fields	Majority Ratio	Target Classes	Records	Augmented Records	New Majority Ratio
Binary	1	stroke	11	8	3	95%	2	5,110	446%	58%
	2	system	16	5	11	95%	2	19,084	594%	56%
	3	champions	11	5	6	96%	2	720	110%	83%
	4	beauty	10	8	2	93%	2	1,260	602%	56%
	5	covid	110	25	32	90%	2	5,644	504%	56%
	6	sylvine	21	1	20	50%	2	5,124	117%	50%
	7	sick	30	14	6	94%	2	3,772	185%	69%
	8	qsar_biodeg	42	12	29	66%	2	1,055	655%	52%
	9	page_blocks	11	1	10	90%	2	5,472	109%	79%
	10	mozilla4	6	2	4	67%	2	15,545	85%	66%
	11	magic_telescope	12	1	10	65%	2	19,020	81%	65%
	12	kr_vs_kp	37	32	0	52%	2	3,196	118%	52%
	13	hmeq_p	15	5	9	80%	2	5,960	607%	54%
	14	compas_two	14	12	2	53%	2	5,278	609%	50%
	15	ada	49	35	6	75%	2	4,147	179%	61%
	16	mimic_icu	49	11	38	86%	2	1,177	182%	66%
	17	titanic	12	6	3	62%	2	891	275%	53%
	18	loan_data	28	4	10	84%	2	19,156	365%	54%
	19	hospital	18	6	12	83%	2	5,956	426%	56%
	20	page_blocks	11	1	10	90%	2	5,472	268%	62%
Multiclass	1	glass	10	1	9	36%	6	214	171%	17%
	2	letter	17	1	16	4%	26	20,000	85%	4%
	3	mfeat_zernike	48	1	47	10%	10	2,000	80%	10%
	4	satimage	37	1	36	24%	6	6,430	114%	17%
	5	soybean	36	36	0	13%	19	683	206%	5%
	6	thyroid	31	15	8	74%	21	9,172	1240%	5%
	7	led	8	8	0	12%	10	500	92%	10%
	8	fetal_health	22	7	14	78%	3	2,126	187%	33%
	9	autos	26	11	14	33%	5	205	129%	20%
	10	ecoli	8	2	5	44%	5	336	173%	22%
	11	segment	11	8	2	67%	4	8,068	213%	25%
	12	sky_server	18	2	12	74%	3	10,000	178%	33%
	13	cirrhosis	20	8	6	56%	4	418	180%	25%
	14	fifa19	18	6	10	12%	25	18,207	237%	4%
	15	body_per	12	2	10	60%	4	13,393	191%	25%
	16	multi_run	17	3	14	64%	5	21,726	258%	20%
	17	microbes	25	1	24	24%	10	30,527	194%	10%
	18	tablet	20	10	10	34%	9	2,000	243%	11%
	19	crystal_structure	18	6	10	61%	5	5,329	244%	20%
	20	accidents	12	3	8	30%	4	10,000	98%	25%
Numeric	1	auto_mpg	9	2	6			398	187%	
	2	bike_sharing	16	6	10			730	173%	
	3	car_price	26	9	14			205	118%	
	4	forest_fires	13	4	8			517	116%	
	5	gemstone	10	3	7			26,967	206%	



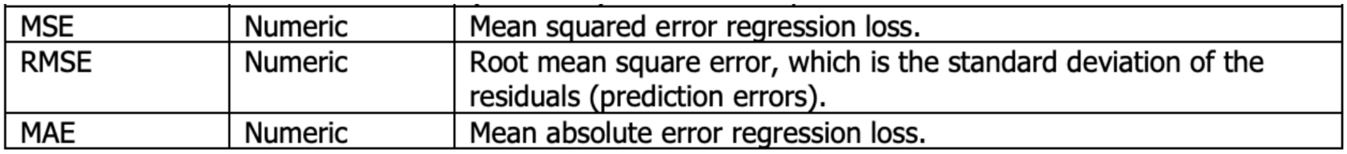
Type of Target	ID	Data	Fields	Categorical Fields	Numeric Fields	Majority Ratio	Target Classes	Records	Augmented Records	New Majority Ratio
	6	jobs	11	5	3			500	242%	
	7	motorcycle	7	2	4			1,061	140%	
	8	profit	5	1	4			50	120%	
	9	real_estate	8	1	7			414	369%	
	10	sales_small	5	1	4			4,572	302%	
	11	student_per	33	29	4			649	163%	
	12	walmart	8	1	7			6,435	272%	
	13	car_seats	12	4	8			400	159%	
	14	insurance	7	4	3			348	190%	
	15	bigmart	12	8	4			8,523	292%	
	16	bodyfat	15	0	15			252	152%	
	17	cellphone	14	5	9			161	246%	
	18	house_rent	12	8	3			4,746	126%	
	19	uso	81	8	73			1,718	242%	
	20	pesticides	13	3	10			8,760	268%	

Table 1: Input Datasets' Metadata

When performing the data enhancement process, the F1 score evaluation metric is used to optimize the free parameters when the data is related to a task with a categorical target, and the RMSE metric is used for the same purpose for data that relates to a task with a numeric target. The metrics used to evaluate and compare the prediction models' performance when trained on the original data versus the optimized data are described in Table 2. For datasets related to tasks with categorical targets (binary or multiclass classification) we measure the difference (improvement) in the following metrics: the F1 score, recall, precision, balanced accuracy, and ROC AUC. So given score A for a model trained on the original training data and score B for a model trained on the optimized training data, we calculate the prediction improvement score C, which is the increased score obtained by the optimized data: C = B – A. For datasets related to tasks with numeric targets, we measure the decrease in the following three metrics: MSE, RMSE, and MAE. In this case, scores A and B represent the models' error, and we calculate the error reduction score C, which is the reduction in error from the original error: C = (A-B)/A.



Evaluation Metric	Target Type	Description
F1 Score	Categorical	Interpreted as the harmonic mean of the precision and recall, where an F1 score ranges between zero (poor performance) and one (best performance). The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is: F1 = 2 * (precision * recall) / (precision + recall). The F1 score is also known as the balanced F-score or F-measure.
Recall	Categorical	The ratio TP / (TP + FN), where TP is the number of true positives, and FN is the number of false negatives; the best value is one, and the worst value is zero. Intuitively, the recall represents the classifier's ability to identify all of the positive samples.
Precision	Categorical	The ratio TP / (TP + FP), where TP is the number of true positives, and FP is the number of false positives; the best value is one, and the worst value is zero. Intuitively, the precision represents the classifier's ability to avoid assigning a positive label to a negative sample.
Balanced Accuracy	Categorical	The average recall obtained by each class, where the best value is one, and the worst value is zero; this metric is suitable for imbalanced datasets.
ROC AUC	Categorical	The area under the curve of the receiver operating characteristic (ROC AUC) based on the prediction scores.



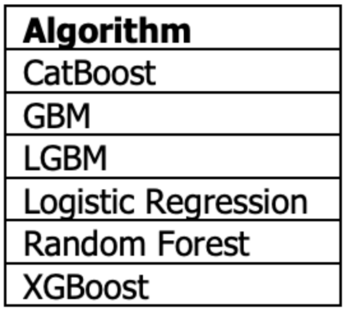
MSE	Numeric	Mean squared error regression loss.
RMSE	Numeric	Root mean square error, which is the standard deviation of the residuals (prediction errors).
MAE	Numeric	Mean absolute error regression loss.

Table 2: The Evaluation Metrics to Evaluate Our Data Enhancement Process

In our evaluation of the data enhancement process, five-fold cross-validation was performed on six prediction models (algorithms). The evaluation flow for a single dataset is as follows:

For each train-test split in the five-fold cross-validation:

- A generative VAE model is trained based on the original training data, and the trained VAE model is then used to generate optimized training data.
- The six algorithms listed in Table 3 are used to:
    - Train model A with the original training data
    - Train model B with the optimized training data



Algorithm
CatBoost
GBM
LGBM
Logistic Regression
Random Forest
XGBoost

Table 3: Algorithms Used to Evaluate the
Data Enhancement Process

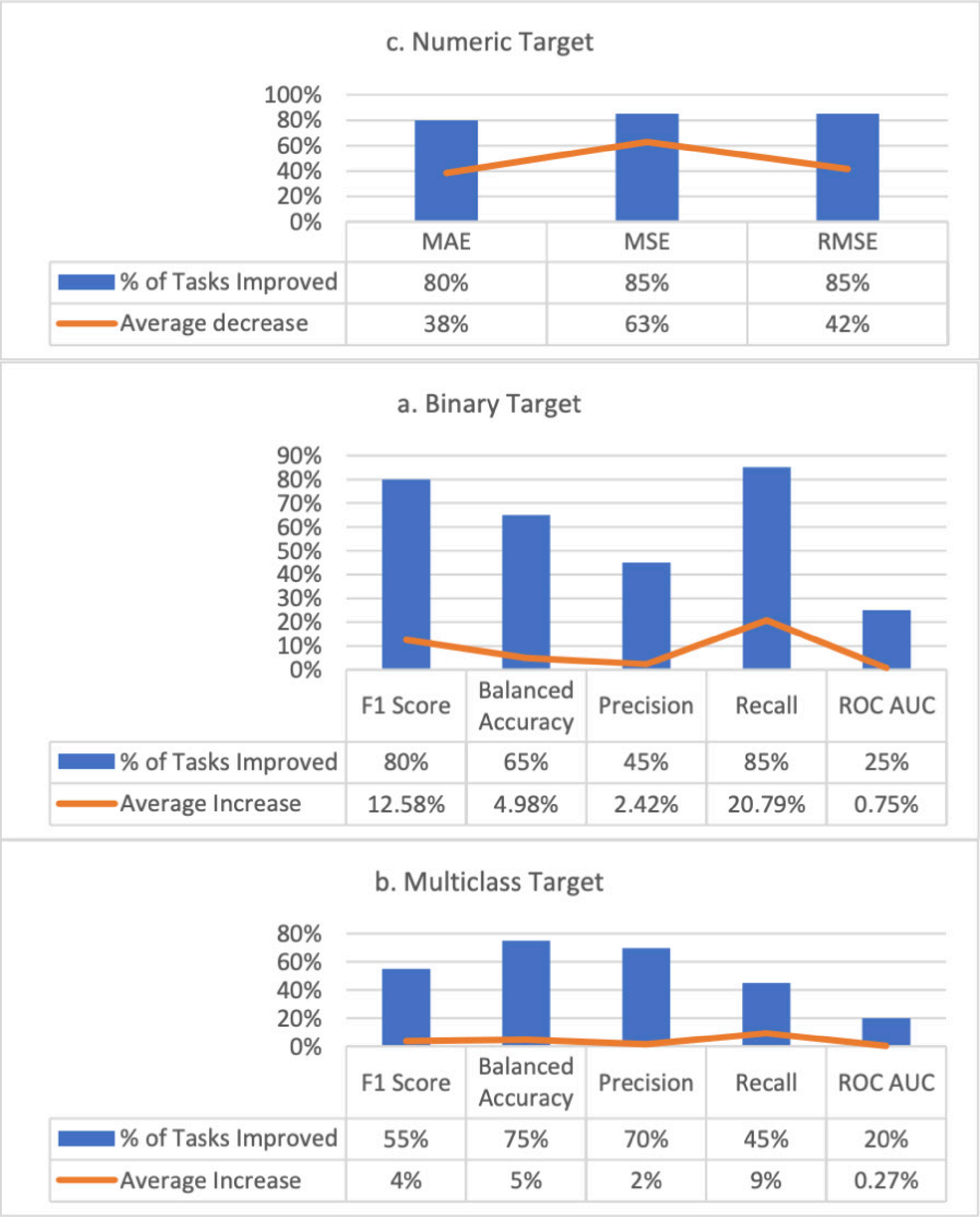
For each of the relevant evaluation metrics:

- The prediction improvement/error reduction score (for categorical/numeric targets respectively) is calculated based on the models' performance on the test data.
- The average prediction improvement/error reduction score is calculated for each algorithm, and later the average prediction improvement/error reduction score for all the examined algorithms is calculated. Finally, the average prediction improvement/error reduction score according to each of the relevant evaluation metrics is calculated for the aggregated scores obtained in the five train-test splits.

Results

Figure 1 provides a summary of the results.

Our results overwhelmingly demonstrate the benefit of optimizing the training data on model performance and the ability of the Datomize-enhanced datasets to produce accurate prediction models. Each of the datasets evaluated focused on a specific type of prediction task (binary target, multiclass target, or numeric target prediction), and in each case, the optimized training data contributed to improved model performance. Note that the results presented represent the average results obtained by all of the algorithms examined based on all five folds. For 80% of the binary target tasks, the F1 score increased by 12.58% when the optimized training data was used; for 85% of these tasks, the recall increased by 20.79%; for 65% of them, the balanced accuracy score increased by 5%; for 45%, the precision increased by 2.42%; and for 25% of the tasks, the ROC AUC increased slightly by 0.75%. For 55% of the multiclass tasks, the F1 score increased by 4% when the optimized training data was used; for 45% of these tasks, the recall increased by 9%; for 75% of them, the balanced accuracy increased by 5%; for 70%, the precision increased by 2%; and for 20% of the tasks, the ROC AUC increased slightly by 0.27%. For 85% of the numeric target tasks, the RMSE decreased by 42% and the MSE decreased by 63% when the optimized training data was used; and for 80% of these tasks, the MAE decreased by 38%.



c. Numeric Target

	MAE	MSE	RMSE
🟦 % of Tasks Improved	80%	85%	85%
🟧 Average decrease	38%	63%	42%

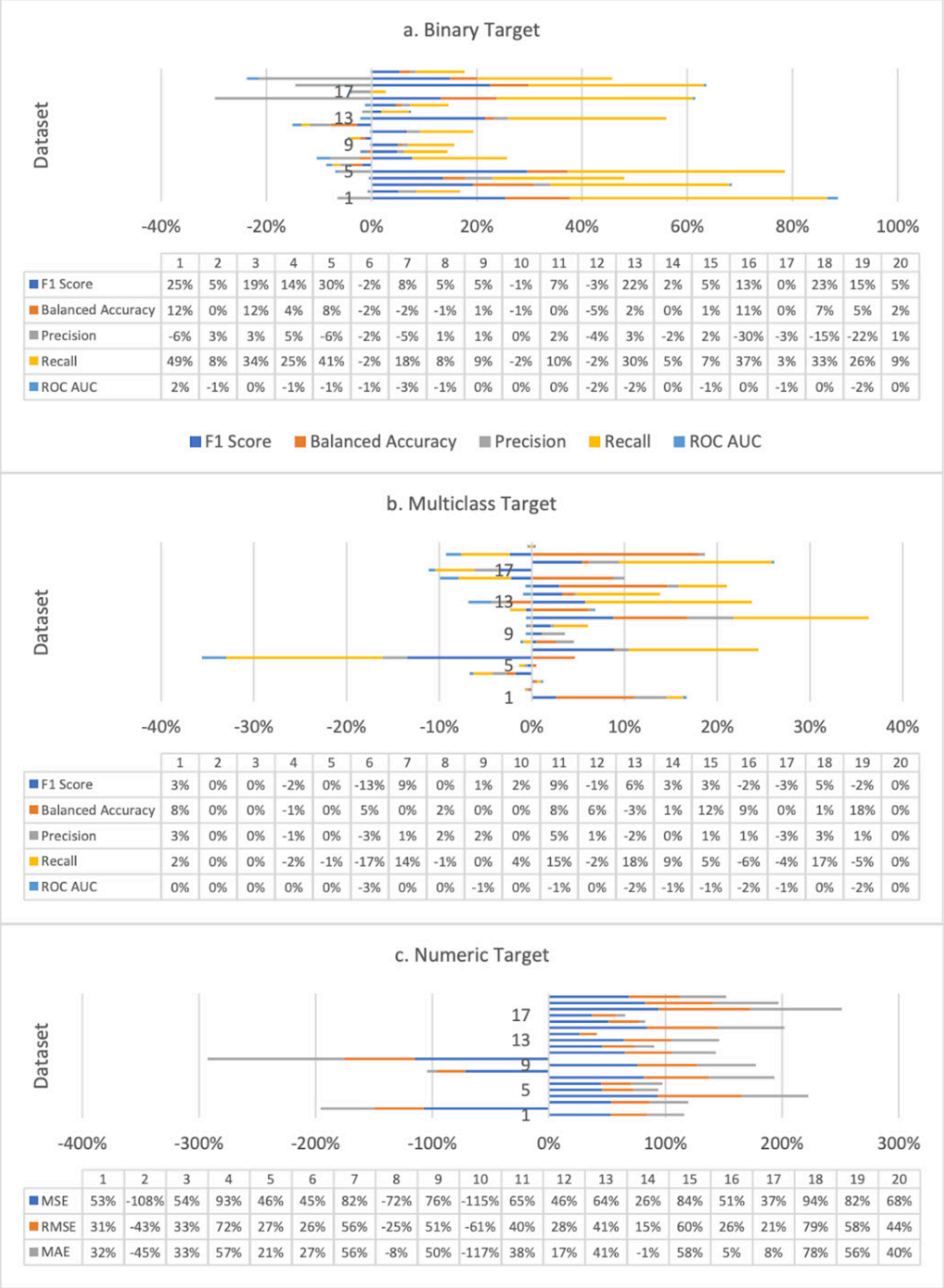
a. Binary Target

	F1 Score	Balanced Accuracy	Precision	Recall	ROC AUC
🟦 % of Tasks Improved	80%	65%	45%	85%	25%
🟧 Average Increase	12.58%	4.98%	2.42%	20.79%	0.75%

b. Multiclass Target

	F1 Score	Balanced Accuracy	Precision	Recall	ROC AUC
🟦 % of Tasks Improved	55%	75%	70%	45%	20%
🟧 Average Increase	4%	5%	2%	9%	0.27%

Figure 1: Average Prediction Improvement After Performing the Data Enhancement Process



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
■ F1 Score	25%	5%	19%	14%	30%	-2%	8%	5%	5%	-1%	7%	-3%	22%	2%	5%	13%	0%	23%	15%	5%
■ Balanced Accuracy	12%	0%	12%	4%	8%	-2%	-2%	-1%	1%	-1%	0%	-5%	2%	0%	1%	11%	0%	7%	5%	2%
■ Precision	-6%	3%	3%	5%	-6%	-2%	-5%	1%	1%	0%	2%	-4%	3%	-2%	2%	-30%	-3%	-15%	-22%	1%
■ Recall	49%	8%	34%	25%	41%	-2%	18%	8%	9%	-2%	10%	-2%	30%	5%	7%	37%	3%	33%	26%	9%
■ ROC AUC	2%	-1%	0%	-1%	-1%	-1%	-3%	-1%	0%	0%	0%	-2%	-2%	0%	-1%	0%	-1%	0%	-2%	0%

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
■ F1 Score	3%	0%	0%	-2%	0%	-13%	9%	0%	1%	2%	9%	-1%	6%	3%	3%	-2%	-3%	5%	-2%	0%
■ Balanced Accuracy	8%	0%	0%	-1%	0%	5%	0%	2%	0%	0%	8%	6%	-3%	1%	12%	9%	0%	1%	18%	0%
■ Precision	3%	0%	0%	-1%	0%	-3%	1%	2%	2%	0%	5%	1%	-2%	0%	1%	1%	-3%	3%	1%	0%
■ Recall	2%	0%	0%	-2%	-1%	-17%	14%	-1%	0%	4%	15%	-2%	18%	9%	5%	-6%	-4%	17%	-5%	0%
■ ROC AUC	0%	0%	0%	0%	0%	-3%	0%	0%	-1%	0%	-1%	0%	-2%	-1%	-1%	-2%	-1%	0%	-2%	0%

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
■ MSE	53%	-108%	54%	93%	46%	45%	82%	-72%	76%	-115%	65%	46%	64%	26%	84%	51%	37%	94%	82%	68%
■ RMSE	31%	-43%	33%	72%	27%	26%	56%	-25%	51%	-61%	40%	28%	41%	15%	60%	26%	21%	79%	58%	44%
■ MAE	32%	-45%	33%	57%	21%	27%	56%	-8%	50%	-117%	38%	17%	41%	-1%	58%	5%	8%	78%	56%	40%

Figure 2: Detailed Results or Each Dataset, Demonstrating the Impact of Our Data Enhancement Process

Conclusions

Our experiments demonstrate how Datomize's data enhancement process significantly improves machine learning model performance. By optimizing the training data, the prediction results dramatically improved when using the exact same models. Improvement was seen across the board, for most of the tasks, for each metric to varying degrees. For example, for 80% of the binary target tasks examined, there was an average increase of 12.6% in the F1 score; for 75% of the multiclass target tasks examined, there was a 5% increase in balanced accuracy; and for 85% of the numeric target tasks, there was an MSE reduction of 63% and an RMSE reduction of 42%. The model-centric approach to AI now serves as the basis of many widely-used open-source models. The recent emergence of the data-centric approach to optimizing AI models has been driven by the ongoing need to improve model performance further and the search for new ways to accomplish this. Generative AI is rapidly rising in prominence as it moves to the public domain where it is taking center stage before a wider audience fascinated with its capabilities. However, for the last few years it has been explored in depth by industry and academic researchers, and generative AI now serves as a source of a wide range of content ranging from images and videos to program code. Given its demonstrated capabilities, it was only natural to consider leveraging generative AI's strengths to improve data quality. And from this coupling, Datomize emerged. Our results highlight the important role that a generative AI approach can fill in a data-centric approach to improving model performance. Datomize's data-centric generative AI-based approach transforms the data we collect into the data we want and need for improved performance.

KEY TAKEAWAYS

• **Evolution of AI Approaches:** The AI landscape has evolved from a model-centric approach, which focused on refining machine learning models, to a data-centric approach, emphasizing the enhancement of data quality.

• **Generative AI's Potential:** The rise of generative AI, especially in structured data, offers promising capabilities. It's not just about generating content but also about improving data quality, making it a pivotal tool in the data-centric approach.

• **Datomize's Data Enhancement Process:** Datomize employs a sophisticated process using a variational autoencoder (VAE) and techniques like crossover and SMOTE. This process optimizes training data, leading to more accurate machine learning models.

• **Significant Performance Improvements:** The results from Datomize's experiments are compelling. For instance, 80% of binary target tasks saw a 12.6% increase in the F1 score, and 85% of numeric target tasks experienced a 42% reduction in RMSE.

• **The Importance of Data Quality:** Data quality directly impacts machine learning predictions. Optimizing data can lead to predictions that are more accurate, efficient, and reliable, reducing risks associated with biases and inaccuracies. Datomize's Role in AI's Future: Datomize's data-centric, generative AI-based approach is poised to redefine how we approach AI. By transforming raw data into the desired format, Datomize is setting a new standard for AI model performance.

Navigating Risks and Rewards: An Intro to Using Generative AI for Data Fabrication

Josh Fourie

Biography

Josh F. is the Chief Technology Officer at Decoded.AI, an Australian startup that provides a SaaS tool to help AI teams align systems with project requirements efficiently.

Josh has been leading Decoded.AI's technology strategy since 2021. He has previous experience managing AI research projects for the Australian Department of Defence and in secure AI.

Before joining Decoded.AI, Josh studied at the Australian National University. He is passionate about leveraging AI responsibly to solve real-world problems. Under his technical leadership, the Decoded. AI platform empowers data teams to rapidly build explanations of their AI systems, enabling better utilization, faster iteration, and a stronger organization-wide data culture.

Josh is an active member of the AI community. He frequently speaks at conferences and meetups on topics like ethical AI design, responsible AI, and explainable AI. In his spare time, Josh enjoys engaging with the latest research and connecting with other AI professionals. He is always on the lookout for companies doing innovative work with AI and exploring potential collaborations.

Navigating Risks and Rewards: An Intro to Using Generative AI for Data Fabrication

By Josh Fourie

Term Key

Reinforcement Learning: A machine-learning technique that trains a model (the "policy") by giving positive and negative rewards for actions taken in a simulated environment.

Generative AI: A model that can 'create' content like audio or images that matches an input prompt such as text, text and an image or a latent space of mathematical variables.

Generator: An Reinforcement Learning policy that can prompt a Generative AI to produce content for which it receives a positive or negative reward that can be used to update the policy.

Generative Adversarial Network: A technique for training generative models that has a Generator which creates content like an image and a discriminator which tries to pick which image is the generated one out of a set of images.

What does it mean to work with 'dataless' AI? Ten years ago, some of us began to get excited about using procedurally generated simulations of digital 'micro-worlds' created on the fly with an element of randomness - to train AI models on challenging tasks. Over time, our capabilities have grown more powerful, and now enable us to fabricate more expansive simulations for more interesting tasks. Generative AI (GenAI) offers us a chance to push the depth of those simulations to unprecedented levels. We anticipate that around 15% of AI companies will rely on these kinds of techniques in the next 5 years, so it is worth considering some of the risks of shifting simulation-building onto GenAI-enabled systems.

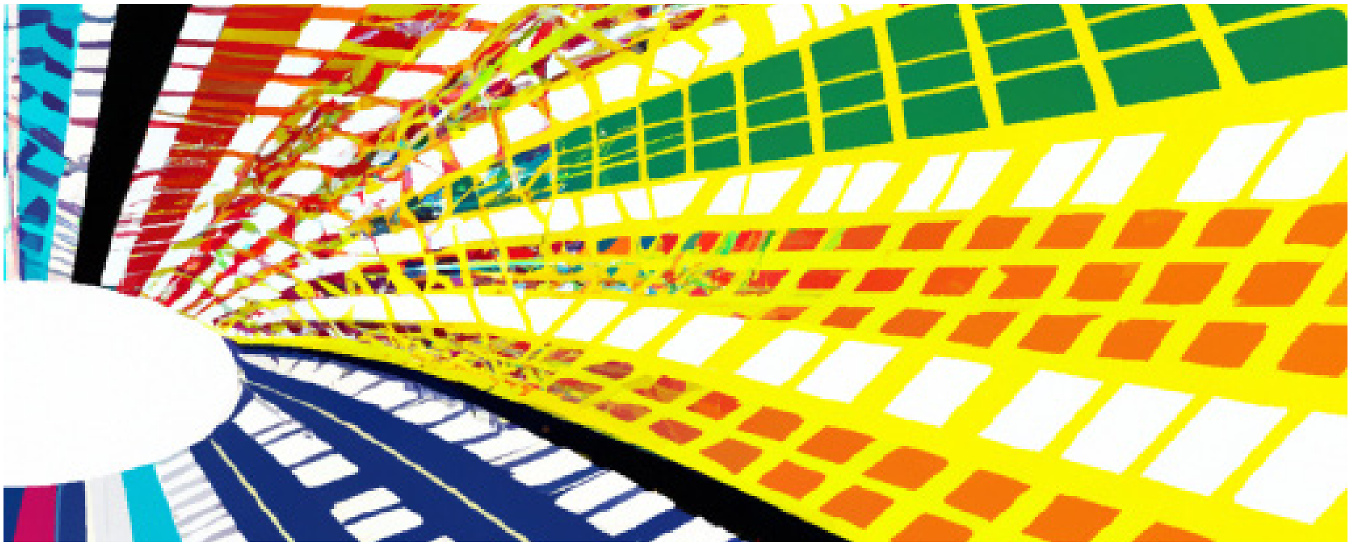


Image generated using DALL-E 2. Prompt: Collecting data in digital ecosystems, pop art style

Breaking Through The Simulation Limits of Reinforcement Learning

One interesting application of GenAl is as a tool for enriching the fabrication of simulated environments in which more sophisticated Al systems are trained, Reinforcement Learning (RL) is a machine-learning technique that imbues an agent with a sense of dynamism, intent and reactivity through repeated interactions with a fabricated simulation. The simulation approach to RL is expensive because developers are constantly embattled with the pain of creating the assets and rules which govern the training of the agent.

GenAI can be used to overcome the asset generation bottleneck of building simulations to enable agents to train more effectively and at scale with less developer time. To push the reactivity and robustness of the fabricated simulation, we 'unfold' it dynamically by training another agent to generate the next interaction ad-hoc based on the progress of the agent. You can think of this system as a school room in which a teacher (the GenAI model) produces content that is set by a director which we call the Generator (RL model #1) to teach a student (RL model #2). Like a school, the content, order and style of the teacher's work shape the student's construction of and value alignment in the world, including emergent bias or appropriateness of heuristics to new and uncontemplated situations.

You can, using this analogy, imagine that a teacher might inadvertently, either by omission or action, create undesirable outcomes or attributes in the student by framing ideas as they are being learned or that a student may draw unexpected or improper lessons from an innocuous lesson. In principle, this is similar to a Generative Adversarial Network (GAN) in which the Generator fabricates episodes of the simulation to nudge the training agent towards a better policy of behaviors.

What we stand to gain is a highly enriched training environment for control tasks like navigation, social tasks like negotiation and management tasks like financial optimisation. However, we risk producing an agent that is unsuitable for a task, that has adopted improper heuristics, or that exposes our users to adverse risk, These risks are exacerbated by the insidious kinds of privacy and bias problems that infest contemporary GenAI models trained on data scraped from the internet. To make matters worse, they are likely to be amplified and drawn out through interactions between the Generator and the training agent.

Navigating the Risks of Simulations Built with GenAI

Our first risk is that the training agent can learn to exploit peculiarities, biases or defects in the fabricated environment to 'outsmart' the reward function and learn a risky policy. This risk exists because RL agents encode an exploratory character in their core algorithm to occasionally make random, counter-intuitive or less-than-optimal choices. We do this so that the agent is more likely to identify useful heuristics in the fabricated environment by experimenting with surprising actions. Often, those surprising actions will yield an unexpected reward and, in practice, it is a common reason that strange defects in the environment are found and exploited. As a result, we are likely to produce an agent which propagates inappropriate bias or improperly 'shortcuts' decisions with heuristics that negatively affect a group of people. To mitigate that risk, developers must invest in writing effective tests that trace the limitations of the system. They must also observe metrics during training that confirm expected behavior rather than relying on visual inspection and debugging during development.

Consider a disaster-response scenario in which an RC-sized car is being trained to navigate a hazardous environment to report the degree of danger to emergency services. It would be useful to maximize the robustness of the simulation by relying on a Generator to procedurally fabricate obstacles, hazards, regional-specific architecture as well as people wearing different clothes, accessories and who are experiencing different reactions. In this case, the assets are meaningful because visual inspection for damage or injury is core to the task. This is a useful paradigm for a developer looking to reduce the cost of the simulation without compromising on the diversity of assets like having to design materials or hazards to place around the city.

We need to be conscious in this scenario about the possibility for bias in our generated assets to impact who is provided with assistance or how those individuals are predicted to behave You can imagine, for example, the training agent learning to prioritize assistance based on clothing or ethnic appearance. This can hap- pen if both agents learn a coded mechanism of communication (a defect) which enables them to cooperate to maximize rewards by marking more rewarding choices with an asset like Clothing.

Alternatively, the Generator might 'inadvertently (independent of the reward function) create scenarios in which buildings that appear with certain religious symbols are more likely to require assistance or associate markers of ethnicity with panicked or less cooperative behavior. These risks are important because the activities, interactions and preponderance that give way to them are baked into the mathematics of both the reward function and the underlying distribution of the GenAI model.

Typically, GenAI models are trained on data scraped from internet sources which can insidiously encode structural, historical and emergent biases as well as include private or proprietary information. Consequently, the Generator is likely to make 'choices' in the contents of the generated assets that reflect and reinforce the cultural paradigms of internet hegemonies. It is also possible for the assets produced for the simulation to resemble genuine people or symbols for which the training agent may develop special heuristics that are 'triggered if those people or symbols are encountered in our physical world (a backdoor). Whilst we stand to gain a lot with this strategy. We also risk encoding our training agent with amplifications of bias and risks that have been encoded into the GenAI model by data scraping practices.

The second risk is that we can be easily distracted by the reality or grandeur of the fabrication from interrogating what an agent is actually learning in a simulation. Simulations are increasingly event-rich and graphically impressive and so we are more likely to rely on things like the intuitive feel of the physics, the visual fidelity of the lighting or the apparent connection of digital objects to physical ones rather than solid risk analysis. As we look to expand the depth of our simulations with GenAI, it is more likely that we will become distracted in our analysis and overlook systems that expose our users to unexpected behavior or failure modes that have historical precedent. It is important that we uncover the history of tools, the people behind them and how they are nudging developers in their analysis. For example, one early, underlying assumption that many people make is that every object, item or experience in the fabricated world can be assigned one unambiguous label that is universally true. Another example is that there is, unfortunately, a direct cost to capturing data in the simulation that can minimize our capacity to retroactively inspect the distribution of experiences during a training run. This means that we are working on the assumption that our metrics about the training are an incomplete picture. Instead, most of our risk analysis will rely on tests confirming the behavior of components of the system. Considerations like these are why, as risk-managers, it is important that we uncover and identify how the tools our teams are relying upon might encourage them to make assumptions or trade-offs in the project. Given the sophistication of modern GenAI, it is easy to see why these risks will become more subtle and harder to detect without necessarily reducing in impact.

Strategies for a Safer Implementation

Systems like these offer a tempting opportunity for well-resourced market actors to boost the quality of their simulations to develop complex, dynamic and reactive AI systems. When thinking about or working with these kinds of systems at a high-level, I recommend taking on one broad philosophy and undertaking at least two kinds of analysis to align the system with the risk tolerance of the creators and users.

As discussed above, it can be easy to think of a system like this as being grounded in the reality of a simulation that will naturally extend into a 'real-world' application (or else the simulation would be useless). This is especially important as we will increasingly see environments that are visually indistinguishable from our experiences of the physical world. As with much of AI, I find it more persuasive to think of these training environments (the 'dataset) as a fabrication; a purposefully crafted leading narrative about a fragment of our world intended to imbue an agent with heuristics that the creator believes are valuable, When managing the risks of a system like this, we need to consistently challenge the narrative that is being told by the fabricated dataset in order to uncover whether there is a credible narrative of risk mitigation.

There are two basic analytical directions that I encourage you to think about. Firstly, when starting to analyze these systems, be concerned with how our tools, paradigms and the constraints of our problem space frame the development of the system and create patterns of risk, Far too often, for example, our fabricated simulations emphasize graphical fidelity, hyper-realistic physics and 'gamification' of tasks (brought on by the reward structure of our training technique). Spend time considering the history of tools, the people behind them and how they are nudging developers in their analysis. For example, one early, underlying assumption that many people make is that every object, item or experience in the fabricated world can be assigned one unambiguous label that is universally true. Another example is that there is, unfortunately, a direct cost to capturing data in the simulation that can minimize our capacity to retroactively inspect the distribution of experiences during a training run. This means that we are working on the assumption that our metrics about the training are an incomplete picture. Instead, most of our risk analysis will rely on tests confirming the behavior of components of the system. Considerations like these are why, as risk-managers, it is important that we uncover and identify how the tools our teams are relying upon might encourage them to make assumptions or trade-offs in the project.

Secondly, focus on the reward functions of the training agent to build an intuition for the system. This might include asking how rewards are determined, what the reward scheme does when certain ideas are in tension, and imagining what is the most adverse subversion of that reward function that one can imagine. Equipped with an intuition for the reward function, you are more likely to be able to imagine experiences or moments of misalignment in the simulation that can help identify failure modes and the most suitable control for that risk.

For example, reward function in the disaster scenario which is tied to the quantity of people identified and rescued is more likely to install an aggressively utilitarian heuristic into the agent and, equipped with that intuition, we can begin to tell more credible narratives of risk in using that kind of function. The idea is that we can use an internal narrative about the incentives of a training environment to build an understanding of the kinds of risks to which we might anticipate a system like this would expose us, our teams or our users, When working with systems with multiple or very complex reward functions it is important to spend time considering how those different forces might intersect to create unexpected outcomes.

Taking Bigger, More Considered Risks in AI

The purpose of the GenAl model is to provide a diversity of assets that would otherwise be too expensive so that the Generator can train a better, faster agent capable of solving complex tasks. Whilst the payoffs are substantial, the risks of relying on a system like this require care to effectively manage because they are produced through the unpredictable interaction of two dynamic systems - at least one of which is trained on data scraped from the internet. To manage those risks, it is critical that we carve out time in our projects to engage in effective testing that defines the limitations of our work as well as understand how our tools, paradigms and constraints encourage us to overlook Al risks.

Over the next few years we are likely to see systems like these gain in popularity as GenAl grows increasingly impressive m graphical fidelity, the use of synthetic data IS normalized and We look to use Al systems to solve more fragile, interactive and expensive problems that require a fabricated training environment. Rather than shying away from these systems, it is important that we seek to understand, mitigate and even wield the risks contained therein so that we can build bolder, more effective and more aligned systems worthy of being called Al.

KEY TAKEAWAYS

• **GenAI and Reinforcement Learning (RL) Synergy:** GenAI can enhance the creation of simulated environments for training RL agents. By using GenAI, developers can overcome the challenges of asset generation, allowing RL agents to train more effectively and at scale, saving developer time. The system can be visualized as a classroom where a GenAI model (teacher) produces content directed by an RL model (director) to train another RL agent (student).

• **Potential Risks with GenAI-Enhanced Simulations:** While GenAI can enrich training environments, it can also introduce risks. These include the potential for the RL agent to learn biases or improper heuristics, especially if the GenAI model has been trained on biased internet data. There's also the risk of the agent exploiting peculiarities in the simulation to achieve rewards in unintended ways.

• **Navigating Biases and Assumptions:** GenAI models, often trained on internet data, can inadvertently introduce biases into the simulation. For instance, in a disaster-response scenario, the agent might prioritize assistance based on clothing or ethnic appearance, or associate certain religious symbols with specific behaviors. Such biases can have real-world implications, especially if the agent is deployed in sensitive tasks.

• **The Illusion of Realism:** As simulations become more graphically impressive and event-rich, there's a risk that developers and users might be distracted by the realism and overlook underlying issues. It's crucial to focus on solid risk analysis and not just the visual or intuitive feel of the simulation.

• **Strategies for Safer Implementation:** To manage the risks of using GenAI in simulations:
Challenge the narrative presented by the simulation to ensure it aligns with real-world expectations.
Understand how tools and paradigms influence the development process and potential biases.
Focus on the reward functions of the training agent to anticipate potential misalignments or failure modes.

• **Future of GenAI in AI Development:** The use of GenAI models promises significant advancements in AI training, especially as graphical fidelity improves and the use of synthetic data becomes more common. However, it's essential to approach these systems with a thorough understanding of their risks and benefits, ensuring that AI systems are both effective and aligned with ethical considerations.

Building the Ethics Stack: Mapping the Ethical AI Innovation Landscape

Abhinav Raghunathan

Biography

Abhinav is currently a Data Scientist specializing in Investment Management Fintech Strategies at Vanguard. He also got his Masters at the University of Pennsylvania majoring in Data Science and focusing on Fair AI / ML. Prior to UPenn, he dual-majored in Computational Engineering and Mathematics at UT Austin, where he also delivered a TEDx Talk on the dangers of algorithmic bias. He publishes independently on topics related to ethical AI and launched EAIDB (eaidb.org), a project meant to provide transparency to the Ethical AI Startup Ecosystem. His writing has appeared on Open Data Science, Towards Data Science, and the Montreal AI Ethics Institute. In the past, he has worked for Vanguard, Starry, FairPlay AI, and Point72 Asset Management.

Building the Ethics Stack: Mapping the Ethical AI Innovation Landscape

By Abhinav Raghunathan

Introduction

The demand for ethical AI services, "responsible AI", has skyrocketed in recent times, in part due to some of the troubling practices employed by large technology companies. Everyday media is full of news of privacy breaches, algorithmic biases, and AI oversights. In the past decade or so, public perception has shifted from a state of general obliviousness to a growing recognition that AI technologies and the massive amounts of data that power them pose very real and immediate threats to privacy, accountability, fairness, and transparency. The Ethical AI Database (EAIDB) seeks to generate another fundamental shift - from awareness of the challenges to education of potential solutions - by spotlighting a nascent and otherwise opaque ecosystem of startups that are actively shifting the arc of AI innovation towards ethical best practices.

The EAIDB is a curated collection of startups that are either actively trying to solve problems that AI and data have created or are building methods to unite AI and society in a safe and responsible manner.

Motivation

We define an "ethical AI company" as one that either provides tools to make existing AI systems ethical or builds products that remediate elements of bias, unfairness, or "unethicalness" in society. The number of such companies has exploded in the last five years to increase the relevance of ethics in AI.

The motivation behind this market research is multidimensional:

- Investors seek to assess AI risk as part of their comprehensive profiling of AI companies. EAIDB provides transparency on the players working to make AI safer and more responsible.
- Internal risk and compliance teams need to operationalize, quantify, and manage AI risk. Identifying a toolset to do so is critical.
- As regulators concretize policy around ethical AI practices, the companies on this list will only grow in salience. They fundamentally provide solutions to the problems AI has created.
- AI should work for everyone, not just one portion of the population. Enforcing fairness and transparency in a black-box algorithms and opaque AI systems is of the utmost importance.

Categories

When we launched EAIDB 2022, we identified five key categories that represented the Ethical AI startup industry and discussed key trends and insights.

1. **Data for AI** – Companies that provide specific services to maintain data privacy, detect data bias early, or provide alternative methods for data collection/generation to avoid bias amplification later in the machine learning lifecycle.
2. **ModelOps, Monitoring, and Explainability** – Companies that provide specific tooling to monitor and detect prediction bias ("quality assurance for ML"), and specialize in black box explainability, continuous distribution monitoring, and multi-metric bias detection.
3. **AI Audits and Governance, Risk, and Compliance (GRC)** – Specialist consulting firms or

platforms that establish accountability/governance, quantify model and/or business risk, or simplify compliance for internal teams within AI systems.

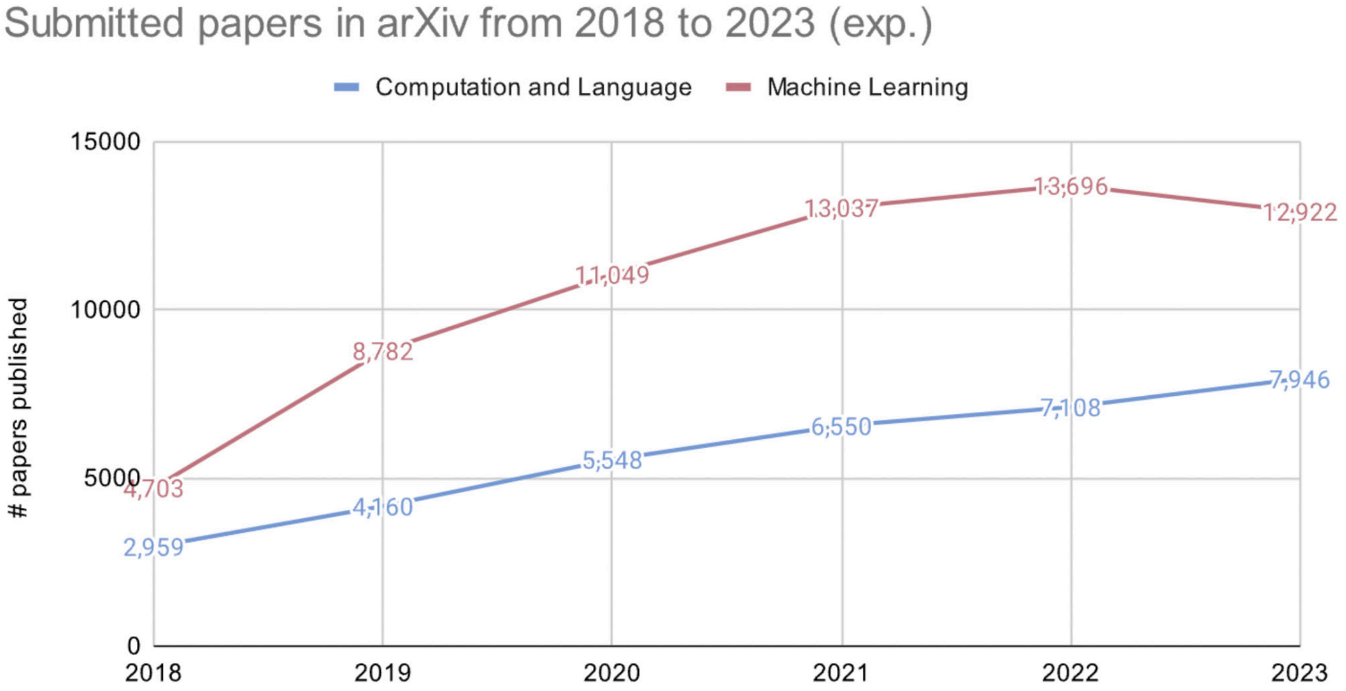
4. Targeted AI Solutions and Technologies – AI companies that attempt to solve a particular ethical issue with a technology that is horizontally-integrated and vertically-applicable. Described as "a more ethical way to __", these companies are usually contained within labels like hiretech, insuretech, fintech, healthtech.

5. Open-Source Solutions – Companies that operate with a focus on ethical principles in AI development and deployment, and they share their work through open-source licenses. This means they released their software, tools, or algorithms to the public so others can use or modify the original code.

Ecosystem Trends

In light of the groundbreaking developments of 2023, including generative AI and other advancements, there have been shifts in the trends of the previously established categories.

<u>The Generative AI Boom</u>

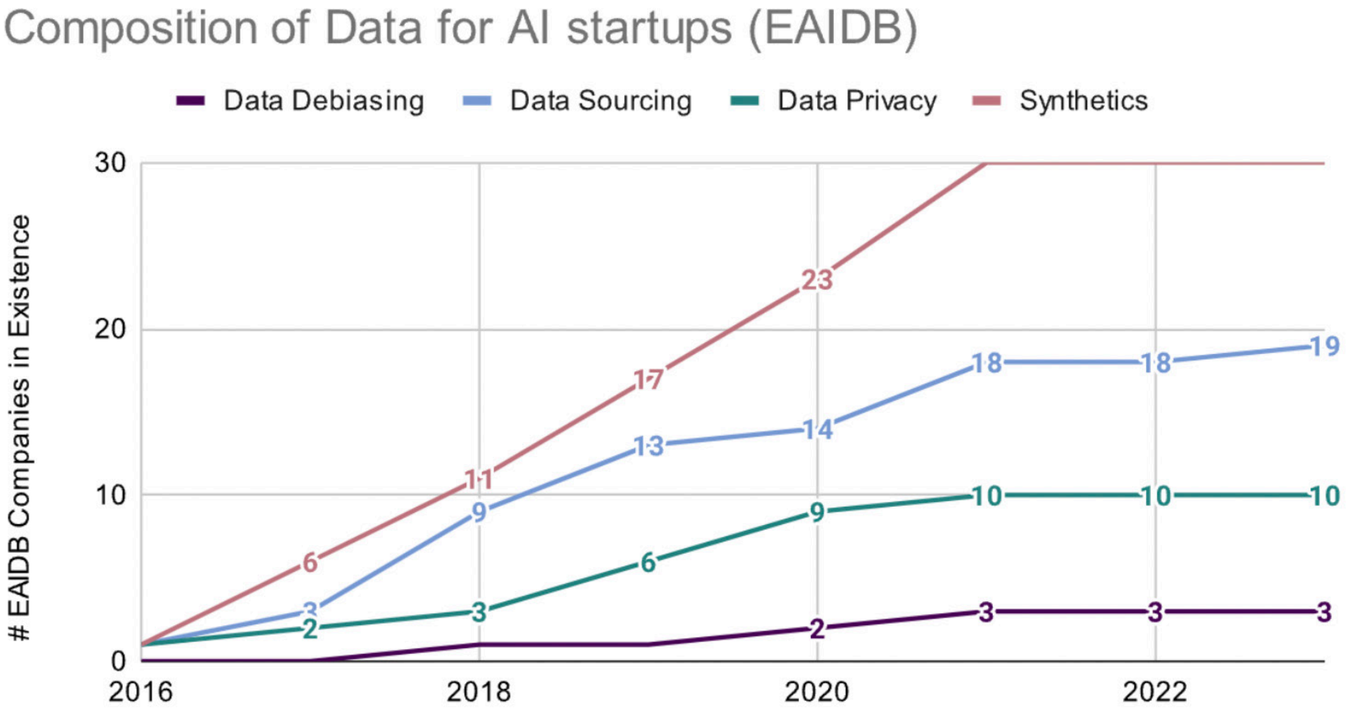


The first half of this year was all about generative AI. The market has benefited greatly from the increased skepticism that providers (particularly OpenAI) have received regarding security, sourcing, environmental cost, toxic outputs, and much more. The truth is, LLMs are the perfect storm of everything wrong with the machine learning process. In response, almost every major provider along the ML value chain has adapted their services to account for some aspect of these downsides (or, at the least, to support LLM development).

We at EAIDB think that, while GenAI is incredibly powerful, it seems reminiscent of Web3 and Crypto in terms of where it is in the hype cycle that seems to power every discussion on the internet. To put the hype in perspective, estimates of generative AI's growth rate are still extremely small compared to other, far more usable and controlled technologies, like causal AI or federated learning (granted, the total market size for these latter technologies is much, much smaller).

Below are some highlighted trends that seem to be driving the industry forward for categories representing a different "type" of ethical AI service. *We've delineated the latest trends and we've introduced several new categories to more accurately capture these emerging patterns.*

Data for AI



Within the Data for AI space, the big theme (intertwined of course with generative AI) is data sourcing and labeling. Not only are LLMs data-hungry, they also perform proportionally with data quality: for organizations building their own LLMs, data quality is often better than data quantity because they allow for smaller, more lightweight models that still perform exceptionally well. In addition, representative data that has been cleared for copyright purposes and purged of toxicity is what is required to build models that can surpass OpenAI's in terms of usefulness.

We've learned a lot from OpenAI's less than ideal methods - from paying Kenyan workers $2/hour for toxicity labeling to copyright infringement, the GPT series is a reiteration of some of the worst aspects of machine learning and artificial intelligence. The importance of responsibility in these instances cannot be overstated.

ModelOps, Monitoring, and Explainability

The MLOps and ModelOps companies in EAIDB have largely reacted positively and have adapted or built aspects of LLMOps to capture some of the use cases in language modeling. There is a great opportunity here simply because the number of new tools, models, papers, and ideas around the topic has skyrocketed in such a way that contrasting various approaches is near impossible.

However, we expect that these solutions will not dominate the market going forward. Just as traditional machine learning turned from an art to a science with low-code platforms, AutoML, and the entire "model and platform builders" space, so too will language modeling become a science.

AI Audits and Governance, Risk, and Compliance (GRC)

AI GRC products cover everything from organization-wide transparency to holistic legal and regulatory compliance to model value and risk assessment. As the sort of "catch-all" at the end of a machine learning pipeline, AI GRC products face consistent competition from literally all sides. One of the most prevalent trends here are the sheer number of companies from other primary business lines revealing new GRC products to build on their existing customer base and provide their users with a more all-encompassing view of their model lifecycles.

Open-Source Solutions

In a loud cry for transparency, there are quite a few startups in EAIDB that offer open source repositories with a "freemium" model: come for the framework and the packages, stay and pay for the scale. HuggingFace did this brilliantly by first establishing a community for open-source models and methods relating to LLMs and adjacent technology, then providing scale via strategic partnerships with Amazon AWS and Microsoft Azure. We've seen more of these come out of the woodwork recently, some of them even raising rounds to continue their mission. It's an interesting business model that leverages the power of crowdsourcing (a formidable force, given that even Google and OpenAI admit their real competition is the open source community!).

We find that there are still large gaps in the open source domain for responsible enablement. The two most abundant types of libraries in this space revolve around MLOps and explainability. There are a few libraries for fairness, bias mitigation, etc., but these are under maintained and understaffed relative to some of the more operational repositories.

Recently, however, there has been a large influx of open source tools released (both by companies in the MLOps space and by independent organizations) for LLM maintenance, development, and validation. Nearly every major responsible MLOps company in EAIDB released packages to assist with LLMOps. In addition, several entities have created open source libraries for regulating LLM behavior and security - Microsoft Guidance, Microsoft Counterfit, Rebuff AI, Guardrails AI are all examples.

Model and Platform Builders

HR model-builders are dying out as other horizontal-oriented companies take over. Finance and healthcare are rapidly scaling.

The decline of new HR-specific startups and algorithms has been repeatedly noted in EAIDB's previous reports. There has not been a single such company on EAIDB's radar targeting the same space since 2020. This may simply be because HR teams are implementing their own ML - as barriers to this technology continue to decrease, the build vs. buy decision leans increasingly towards "build."

Finance is an almost opposite story. There has not been much momentum in the finance vertical for responsible AI enablement just yet, but there are certainly relevant catalysts that are pushing the industry towards adopting better AI, for example from Consumer Financial Protection Bureau's report to the US Congress in 2022.

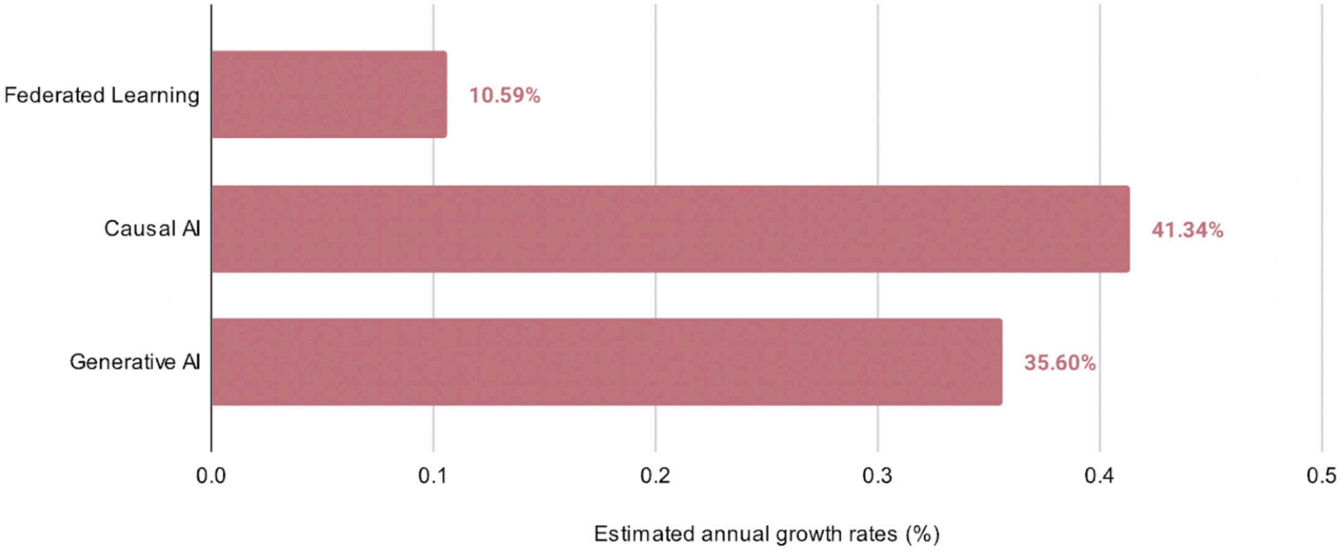
Alternative Machine Learning

On the Alternative ML side, there are really three frameworks that stand out against the backdrop of centralized ML: causal AI, federated learning, and neurosymbolic AI (not as applicable).

Causal AI: Causal AI is growing at a much faster rate than generative AI. This is potentially because of its essentially unlimited use cases in healthcare and other experiment-driven fields where generative AI is just not as useful. There are discussions, however, of uniting the two - generative AI models with the ability to make causal inference.

Federated AI: Still a growing market, federated learning solves problems associated with high data transfer costs and sensitive data usage by allowing machine learning algorithms to be constructed in a distributed fashion such that data never leaves its home. However, there is also additional interest in Large Language Models (LLM) trained in a federated environment.

Effective growth rates of emerging AI technologies (MarketsandMarkets)



<u>AI Security</u>
AI Security consists of “security for AI,” not “AI for security.” Startups in this area address some of the more dangerous aspects of models in production. This category also comprises the greatest amount of current investor attention because it is the most approachable and understandable from a business value perspective.

Companies in this space offer model testing solutions (like adversarial testing, inversion testing, penetration testing, prompt injection, etc.) and sometimes couple them with dashboards or model inventories (akin to some products in the GRC space).

Those that provide a more holistic view of an organization’s model attack surface target CISOs or other C-level executives. Others have strong proprietary technology meant to thoroughly test and secure AI pipelines.

Conclusion

Moving from these changes in ethical AI trends of 2023, it’s evident AI development has played a pivotal role in shaping the global sphere of AI as well as funding patterns of VCs. As we conclude, it’s crucial to recognize the transformative influence of generative AI in driving these trends, underscoring its profound impact on the broader landscape of technological evolution.

<u>Global Demand</u>
Demand is increasing, steadily. The EU will be the first to pass an all-encompassing AI Act that will surely do for the responsible AI market what GDPR did for privacy. Other companies like China, India, and Australia are following suit as AI rapidly approaches maturity.

There is still limited incentive for businesses that are not in highly regulated environments because there is no demonstrated or tangible value of responsible AI enablement. This is changing, however, in part due to generative AI. Generative AI has shown most people in technology what the true impact of concepts like hallucination, misinformation, toxicity, illegally acquired or biased data, and lack of visibility is. Businesses want to harness the power of this technology without subjecting themselves to the aforementioned additional risks. This takes some level of investment - whether it is a build or buy decision. Large enterprises will most likely build their own solutions, whereas medium-to-large enterprises may choose to buy. Smaller organizations globally are still identifying their pain points.

Funding Patterns

There were a lot of new funding rounds raised in 1H2023. The VC market has cooled substantially since 2021 due to limited capital availability and surprising financial events like SVB's collapse. The hype around GenAI hasn't actually helped dealmaking, either. It is quite clear that most of the market does not really understand GenAI companies and their technology.

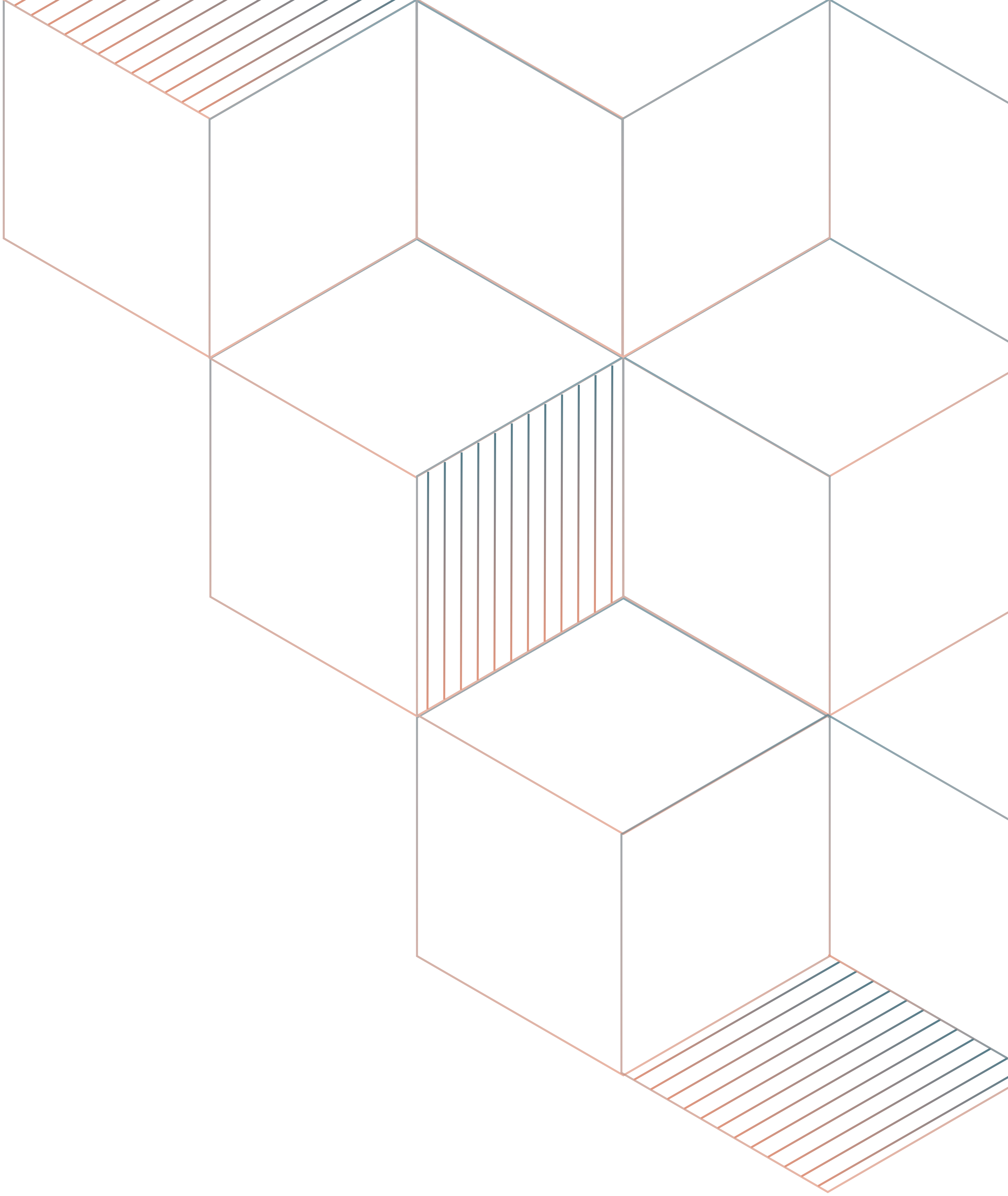
Interestingly, the number of funding rounds within EAIDB has stayed relatively on pace with 2022. The total rounds raised in the first half of 2023 is comparable to last year's despite these headwinds.

This is a nascent ecosystem but it is growing rapidly and is expected to increase in momentum as motivations improve. Incipient measures to track and measure this area will increase in turn. EAIDB reports will be published on a quarterly basis to lift the veil and spotlight both the importance and growth of this space. Over time, trend lines will emerge and taxonomies will shift to adapt to the dynamic reality of this ecosystem.

KEY TAKEAWAYS

Generative AI's Dominance and Challenges:

- Rise of Generative AI: The first half of 2023 saw a significant boom in generative AI, with its growth and influence becoming a major talking point. Despite its promise, generative AI faced skepticism regarding security, sourcing, environmental costs, and potential for toxic outputs. While generative AI is surrounded by hype, its growth rate is still smaller compared to other technologies like causal AI or federated learning.
- **Emerging Trends in Ethical AI:**
  - Data for AI: Emphasis on data sourcing and labeling, with a focus on quality over quantity and the importance of copyright and toxicity considerations.
  - Model Operations and Monitoring: The rise of tools and methodologies for managing and monitoring large language models.
  - AI Governance and Compliance: A surge in products and solutions that ensure transparency, legal compliance, and risk assessment in AI deployments.
  - Open-Source Movement: A notable trend towards open-source solutions, with companies offering frameworks and tools that cater to the community's demand for transparency.
- **Market Dynamics and Future Outlook:**
  - Global Demand: Anticipation of regulatory changes, especially in the EU, that could shape the responsible AI market similarly to how GDPR influenced privacy.
  - Funding Patterns: Despite challenges like the cooling VC market and limited capital availability, funding rounds in 1H 2023 remained consistent with 2022.
  - Ecosystem Growth: The ethical AI ecosystem, though in its early stages, is growing rapidly, with expectations of increased momentum and more comprehensive tracking and reporting in the future.



INVESTING IN AN ETHICAL FUTURE: BUILDING AI RESPONSIBLY

Special: Harnessing the Power of Generative AI in Cybersecurity

Anik Bose

Biography

Anik Bose is the executive director of EAIGG, and is a General Partner at Benhamou Global Ventures. He has over 25 years of active venture capital and corporate development experience, with particular emphasis on transaction structuring and strategic planning. This experience includes seven years as SVP, Corporate Development at 3Com Corporation and eighteen years at BGV. In his role at SVP, Corporate Development at 3Com he played a significant role in company strategy, managed a $250MM corporate venture fund, spearheaded 20 venture capital investments, and executed numerous spin-out transactions, as well as several large M&A transactions totalling over $1.4 Billion in value. Bose has extensive global experience as a private company director, having served on the board of several 3Com Ventures, BGV portfolio companies and H3C. Bose holds a BA in Economics from the University of Delhi, India, as well as an MBA from Boston College.

Harnessing Generative AI in Cybersecurity

Alberto Yépez

Biography

Alberto Yépez is a Co-Founder and Managing Director at Forgepoint Capital. He is a serial entrepreneur, operator, and investor known for his expertise in cybersecurity and digital infrastructure software. At Forgepoint, Alberto led investments in AttivoNetworks (acq. SentinelOne), Area 1 (acq. Cloudflare), BehavioSec (acq. RELX). He currently serves on the boards of Constella Intelligence, CyberCube, Huntress, NowSecure, ReversingLabs, and Uptycs. Before Forgepoint, he led cybersecurity investments at Trident Capital and held leadership roles as Founder, Chairman and CEO of enCommerce (acq. Entrust), then President of Entrust (acq. Thoma Bravo), and Chairman and CEO of Thor Technologies (acq. Oracle). Alberto's diverse experience also includes tenures as an Entrepreneur in Residence at Warburg Pincus, a venture consultant at Bain Capital, and a consultant for the U.S. Department of Defense. Alberto is affiliated with the Aspen Institute's Global Cybersecurity Group and the Hispanic IT Executive Council (HITEC). He has been a board member for the NVCA, the University of San Francisco, and the World Economic Forum's Technology Pioneers Community. He was recently recognized by Fortune as a top VC Dominating Cybersecurity Investing and on TAG Cyber's "Fifty to Watch: People Shaping Cyber" list.

Harnessing Generative AI in Cybersecurity

By Anik Bose and Alberto Yépez

Artificial Intelligence is currently experiencing its “Netscape” moment, propelled by the advent of potent Generative AI models such as Chat GPT. Research conducted by McKinsey estimates that generative AI could contribute an equivalent of $2.6 trillion to $4.4 trillion annually to the global economy. (To put this into perspective, the United Kingdom’s total GDP in 2021 was approximately $3.1 trillion.) According to their analysis, about 75% of the potential value generative AI use cases could deliver is concentrated in four areas: customer operations, marketing and sales, software engineering, and R&D across industries. Unsurprisingly, AI is dominating conversations across the cyber world as businesses rapidly adopt and develop AI-based technologies- and/or react to their sudden rise and accessibility. So what are the implications on AI and Cybersecurity?

AI and Generative AI: Context and Definitions

Let’s begin with our context. AI is hardly new despite the intense hype cycle we find ourselves within. AI was first defined as an academic discipline in the mid-1950’s and has since gone through its own boom and busts – periods of intense interest (and funding) followed by “AI winters” and so on. Before the advent of Generative AI, our understanding of AI’s impact on cybersecurity was twofold. First, we recognized the application of AI for protection and detection, either as part of new solutions or as a means to bolster more conventional countermeasures. Second, we acknowledged the necessity to secure AI itself- both as a protective technology and as a tool used by threat actors to develop new attack vectors. Use cases varied from Transaction Fraud Detection, Botnet detection, File-based Malware detection, Network risk assessment, Vulnerability remediation, user authentication, endpoint protection (XDR), and spam filtering.

Today, with the release of several Generative AI platforms, we anticipate the Cybersecurity sector to be profoundly impacted in additional ways including:

1. Amplifying the capabilities of malevolent actors through attack vectors such as evasion, extraction, and enumeration attacks.
2. Bridging the cyber skills gap with powerful AI assistants, to boost the productivity of enterprise cyber teams. These include those launched by incumbents like Crowdstrike and Microsoft.
3. Elevating compliance guardrails around data privacy and output data verification to ensure responsible AI deployment.

Before delving deeper, it’s essential to clarify a few key definitions:
1. AGI (Artificial General Intelligence): AGI refers to highly autonomous systems that can outperform humans at most economically valuable work. AGI encompasses general intelligence and is capable of understanding, learning, and applying knowledge across a wide range of tasks. The goal is to replicate human-level intelligence, with the potential to exhibit self-awareness and consciousness. Our hypothesis is that Threat Intelligence Platforms (TIP) will shift towards GPT-like chats as a more effective information source for users, either as auto prompts and API feeds based on detection Indicators of Compromise (IOCs), or interactive for R&D, similar to how Microsoft Copilot is used for app development, Security, and M365, and GitHub Copilot is used for programming.
2. GPT (Generative Pre-trained Transformer): GPT is a specific type of AI model developed by OpenAI (for clarity, the popular ChatGPT is an AI chatbot app powered by GPT, similar to how a Lenovo or

Dell laptop might be powered by Intel). Models such as GPT-3 and GPT-4 are designed for language generation tasks. They are pre-trained on large volumes of text data and can generate human-like responses given a prompt. These models excel at tasks like natural language understanding, text completion, and language translation. Our hypothesis is that AGI will improve interpretive systems (SOAR and Anti-Fraud) as Large Language Models (LLMs) and Small Language Models (SLMs) are harnessed for their most suitable functions.

New Attack Vectors: Enhancing the Capabilities of Malevolent Actors

Generative AI is a double-edged sword. While it holds immense potential for improving cybersecurity defenses, it also amplifies the capabilities of malevolent actors. By exploiting the capabilities of sophisticated AI models, attackers can devise new attack vectors that traditional security measures may struggle to counter:

1. **Evasion Attacks:** In evasion attacks, the adversary uses generative AI to create inputs that are designed to be misclassified by AI-based detection systems. For example, they could manipulate malware so it appears benign to the security system, thereby evading detection. Generative AI, with its ability to understand and generate data patterns, can significantly improve the success rate of these evasion attempts.

2. **Extraction Attacks:** Extraction attacks refer to scenarios where an adversary trains a model to extract sensitive information from a system, leading to potential data breaches. The advent of Generative AI means that attackers can train models to mimic the behavior of legitimate users or systems, thus tricking security measures and gaining unauthorized access.

3. **Enumeration Attacks**: Enumeration attacks involve using generative AI to discover system vulnerabilities. Hackers can automate the process of testing different attack vectors, rapidly identifying weak points in a system that they can then exploit.

4. **Influence Attacks on Classifiers**: Influence campaigns have been demonstrated in social media and securities/commodities trading systems' reliance on AI repeatedly over the past decade or more – including election cycle and quarantine era mis/disinformation as well as the manipulation of market pricing and performance news. As generative AI is used for more specific, yet broader contexts and concepts in organizational functions, those same techniques will be exercised to exploit the dependencies on knowledge offered to organizations and consumers.

5. **Poisoning Attacks on Data:** One simple example is in Copilot and generative AI code samples that hallucinate functions or resources that hackers may take advantage of to create malicious resources that are subsequently called by that code. This vulnerability requires code validation and testing before production release, which is generally a common activity in modern CI/CD development. This means that even development systems can be compromised and offer back doors for more nefarious software supply chain compromises, especially since those development systems are rarely subject to network isolation or security controls levied on production systems.

As Generative AI continues to evolve, we anticipate an increase in these types of sophisticated attacks. Therefore, it is imperative for both incumbent and startup entities in the cybersecurity sector to remain vigilant and proactive, developing countermeasures that anticipate these new forms of threats.

While this may seem daunting, we believe it is also an opportunity for cybersecurity innovation. The challenges posed by generative AI-powered cyberattacks necessitate novel solutions, opening

Malevolent actors can be enabled further because of generative AI capabilities. https://venturebeat.com/security/10-ways-chatgpt-and-generative-ai-can-strengthen-zero-trust/

new frontiers in the cyber defense sector. Our discussions with key industry players reveal a robust willingness and preparedness to address these concerns.

Broad Yet Precise: Generative AI's Impact on Cybersecurity Innovation

Generative AI has significant potential to influence cybersecurity innovation, both in established companies (incumbents) and startups. Here's how generative AI is shaping cybersecurity:

1. **Anomaly Detection and Analysis:** Generative AI models, trained on substantial datasets of known malware and cyber threats, can identify patterns and generate new threat signatures. This aids real-time threat detection and analysis, empowering security systems to proactively identify and respond to emerging threats. Generative AI models are used to detect adversarial attacks, where bad actors attempt to manipulate or deceive AI systems.
2. **Security Testing and Vulnerability Assessment:** Generative AI can automate security testing by generating and executing various attack scenarios to identify vulnerabilities in software, networks, or systems.
3. **Password and Credential Security:** Startups are using generative AI to develop password and credential security solutions.
4. **Malware Generation and Defense:** Generative AI can be employed to generate new malware samples for research purposes and to strengthen antivirus and anti-malware systems.
5. **Security Operations Automation:** Generative AI models can automate routine security operations while augmenting SOC analyst productivity.

The Need for Guardrails: The Generative AI Accuracy Problem

Generative AI has its limitations - primarily around consistently providing accurate outputs. Therefore, what guardrails are needed to reduce risks and ensure success with broader adoption? Generative AI tools like ChatGPT can augment subject matter experts by automating repetitive tasks. However, they are unlikely to displace experts entirely in B2B use cases due to AI's lack of domain-specific contextual knowledge and the need for trust and verification of underlying data sets. Broader adoption of Generative AI will stimulate an increased demand for authenticated, verifiable data, free of AI hallucinations. This appetite will spur advancements in data integrity and verification solutions, alongside a number of other ethical AI issues such as privacy, fairness, and governance innovations. Boards of Directors now more vocally demand the responsible use of AI to improve operational efficiency, customer satisfaction and innovation, while safeguarding customer, employee and supplier data and protecting intellectual property assets.

On Near-Term Innovation: Incumbents' Edge

Incumbents carry the advantage of pre-existing infrastructure, high-compute resources, and access to substantial datasets. Consequently, we anticipate a surge of innovation from these entities in the near term. Industry stalwarts such as Crowdstrike, Palo Alto Networks, Microsoft, Google, IBM and Oracle are already harnessing Generative AI to bolster their security solutions. Here's an exploration of their endeavors:

Crowdstrike:

- Threat Detection and Response: Crowdstrike employs generative AI to detect and respond to advanced threats in real-time. Their AI-integrated platform, Falcon, scrutinizes large amounts of data to discern patterns and threat indicators, enabling swift detection and response to cyber threats.
- Adversarial Attack Detection: Utilizing generative AI models, Crowdstrike can detect and counter adversarial attacks like fileless malware and ransomware. Their AI algorithms are capable of pinpointing suspicious behavior, anomalies, and threat indicators.
- AI-Driven Security Analytics: By leveraging generative AI, Crowdstrike enhances its security analytics capabilities, thereby enabling the identification of intricate attack patterns, threat prediction, and the generation of actionable insights for security teams.

Palo Alto Networks:

- Threat Intelligence and Automation: The company integrates generative AI into their security platform, Cortex XSOAR, automating threat intelligence and incident response processes. Their AI algorithms sift through extensive threat data, equipping security teams with actionable insights and automated playbooks for efficient threat response.
- Malware Analysis: Generative AI models power advanced malware analysis. This helps companies understand emerging threats, devise effective countermeasures, and fortify cybersecurity solutions.
- Behavioral Analytics: Generative AI aids in developing behavioral analytics models that learn standard user, device, and network behaviors to detect anomalies and potential security breaches.
- Security Policy Optimization: By using generative AI, Palo Alto Networks optimizes security policies through the analysis of network traffic patterns, user behavior, and threat intelligence data, dynamically adjusting security policies for robust protection against emerging threats.

Microsoft:

- SOC Automation: MS's Security Copilot is a large language AI model powered by OpenAI's GPT-4, combined with a Microsoft security-specific model that incorporates what Microsoft describes as a growing set of security-specific skills informed by its global threat intelligence and vast signals volume. Security Copilot integrates with the Microsoft Security products portfolio, which means that it offers the most value to those with a significant investment in the Microsoft security portfolio.
- Human-in-the-Loop Augmentation – While Security Copilot calls upon its existing security skills to respond, it also learns new skills thanks to the learning system with which the security-

specific model has been equipped. Users can save prompts into a “Promptbook,” a set of steps or automations that users have developed. This introduction is likely to be resonant and disruptive because of the human aspect that remains — and will remain — so vital to security operations. The ability of large language AI models to comb through vast amounts of information and present it conversationally addresses one of the primary use cases of automation in SecOps: gathering the context of incidents and events to help analysts triage and escalate those that pose a significant threat.

Google:

- Vulnerability and Malware Detection: Google announced the release of Cloud Security AI Workbench powered by a specialized “security” AI language model called Sec-PaLM. An offshoot of Google’s PaLMmodel, Sec-PaLM is “fine-tuned for security use cases,” Google says — incorporating security intelligence such as research on software vulnerabilities, malware, threat indicators and behavioral threat actor profiles.
- Threat Intelligence: Cloud Security AI Workbench also spans a range of new AI-powered tools, like Mandiant’s Threat Intelligence AI, which will leverage Sec-PaLM to find, summarize and act on security threats. VirusTotal, another Google property, will use Sec-PaLM to help subscribers analyze and explain the behavior of malicious scripts.

IBM:

- Threat Detection and Response: IBM’s QRadar Suite is a subscription-based (SaaS) offering that combines AI-enhanced versions of IBM’s existing threat detection and response solutions into a comprehensive global product. The new QRadar Suite goes beyond traditional security information and event management (SIEM) capabilities, aiming to provide a unified experience for security management. Its goal is to assist organizations in managing extended detection and response (EDR/XDR) capabilities, SIEM functionalities, and Security Orchestration Automation and Response (SOAR) in cybersecurity.
- Security Compliance: IBM’s approach to security and compliance in highly regulated industries, such as financial services, emphasizes the importance of continuous compliance within a cloud environment. By integrating the Security and Compliance Center, organizations can minimize the risks associated with historically challenging and manual compliance processes. The solution enables the integration of daily, automatic compliance checks into the development lifecycle, ensuring adherence to industry standards and protecting customer and application data.

Oracle, SAP, Salesforce and other enterprise application providers are beginning to provide comprehensive AI service portfolios integrating their cloud applications and their existing AI infrastructure with state-of-the-art generative innovations. Their unique approach and differentiation means their customers will have complete control and ownership of their own data inside their “wall gardens” to derive insights and avoid data loss and contamination.

The incumbents not only have the company and customer install base and diverse platform to develop, test, and secure the safe and productive use of Generative AI / AI in general – but also having their own first party security products (Google’s Mandiant and Microsoft Security/Sentinel along with IBM’s Q Labs and Resilient acquisitions) that are using generative AI to power automated threat intel and security…while needing to retain human in the loop decision-making throughout the SDLC (and modern SOCs).

Longer Term Innovation: Advantage Startups

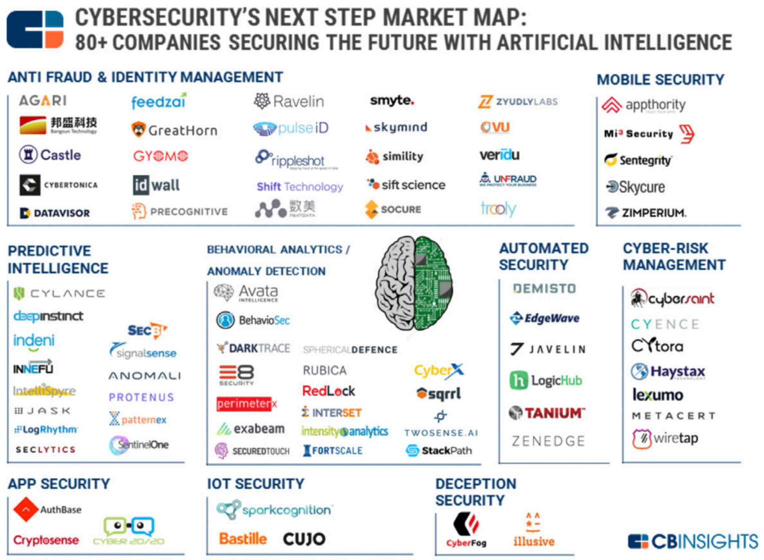
Startups offer innovative, agile solutions in the realm of generative AI for cybersecurity. However, the investment climate for generative AI-driven cyber solutions is still nascent, given the limited number of attacks witnessed to date involving the AI attack surface.

The pivotal role of data cannot be overstated. For startups to flourish, they must leverage open-source LLMs while enriching data with proprietary information. We anticipate that synthetic data innovation and Robotic Process Automation (RPA) will play crucial roles, especially in regulated sectors like financial services and healthcare that have unique data privacy requirements. However, synthetic data is not expected to significantly influence decision support automation, such as privileged access management.

Another key area for startup innovation exists around Verification and Testing, driven by mounting enterprise demand to harness Large Language Models (LLMs). Other noteworthy areas of opportunity include Explainability, ModelOps, Data Privacy for Generative AI applications, Adversarial AI/Data Poisoning, Autonomous Security Operations Centers (SOCs), Differential Data Privacy, and Fraud Detection.

Capital efficient startups will need to utilize existing infrastructure (foundational models) and concentrate on applications that add value through Single Language Models (SLM) via contextual data enrichment. Acquiring proprietary datasets may also be a strategic move for startups aiming to establish a competitive edge.

Furthermore, we posit that the compliance and regulatory environment shaped by the EU Act will direct startup innovation toward responsible AI and Governance, Risk Management, and Compliance (GRC). Notably, the founder DNA in this space will require a unique blend of cybersecurity domain expertise paired with generative AI technical prowess.



Cybersecurity startup industry and market
https://www.cbinsights.com/research/cybersecurity-artificial-intelligence-startups-market-map/

In Conclusion

We anticipate strong innovation at the intersection of Cybersecurity and Generative AI, fueled by incumbents in the near term and startups in the long term. Automating repetitive tasks with Security Co-pilots will go a long way towards addressing the cyber skills gap, while newfound protection and defense capabilities enabled by Generative AI will help secure large enterprise datasets and enable more effective identity orchestration to prevent breaches amid expanding attack surfaces. Morgan Stanley predicts that Cybersecurity is ripe for AI automation representing a $30Bn market opportunity. The bar on compliance guardrails will be raised in this space given the ethical concerns around the accuracy of Generative AI outputs (hallucinations), increasing the need for human-in-the-loop, regulations and raising the stakes to build an “ethics stack” to complement and safeguard the explosive AI technology stack. Finally, enterprise CTA’s (committees of technology and architecture) will increasingly need to embrace responsible application of Generative AI to succeed and compete.

Board of Directors will play an important role to demand good governance and the use of responsible AI, while protecting the key information assets of every business.

KEY TAKEAWAYS

- **Generative AI's Economic and Historical Context:**
  - Experiencing a significant moment in its development and adoption.
  - Potential to contribute $2.6 trillion to $4.4 trillion annually to the global economy.
  - AI's historical evolution from the 1950s, with periods of interest and "AI winters."

- **Dual Impact of Generative AI on Cybersecurity:**
  - Enhances capabilities of malicious actors, introducing new threats like evasion and extraction attacks.
  - Offers opportunities to bolster cybersecurity defenses and develop novel solutions.

- **Generative AI's Influence on Cybersecurity Innovation:**
  - Established companies, including Crowdstrike and Microsoft, are harnessing Generative AI to enhance security solutions.
  - Startups focus on areas like verification, testing, and data privacy, leveraging open-source models and enriching data.

- **Future Outlook and Ethical Concerns:**
  - Morgan Stanley predicts a $30Bn market opportunity in AI automation for cybersecurity.
  - Rise of Generative AI brings ethical concerns around accuracy, necessitating human oversight, regulations, and an "ethics stack."
  - Boards of Directors will play a crucial role in ensuring good governance and responsible AI use.

Four Investors Explain Why AI Ethics Cannot Be an Afterthought

Alexis Alston

Biography

Alexis Alston is currently the Principal at Lightship Capital, where she champions seed-stage founders in the Midwest and South. She also holds a pivotal role as the Director of Corporate Partnerships at Lightship Foundation, spearheading partnerships for Black Tech Week and other strategic programs. Alexis earned her Bachelor's degree in Business from Brown University, with a focus on entrepreneurship and chemistry. Before her tenure at Lightship, she was an Analyst at Advantage Capital, a firm dedicated to financing businesses in underserved communities. Alexis's commitment to inclusivity and community is further highlighted by her board membership at the Special Needs Support Center, advocating for families with special needs in the Upper Valley. In her earlier career, she gained experience in eCommerce analytics at Homesite Insurance and investment research at Chatham Capital.

Four Investors Explain Why AI Ethics Cannot Be an Afterthought

Justyn Hornor

Biography

Justyn Hornor is currently the Chief Technology Officer at Tattd, Inc., where he spearheads the tech stack, AI-driven capabilities, and go-to-market strategies. He also serves as the Founder and Managing Partner of Rocket Product Launch Consulting, focusing on market validation and innovation for diverse sectors. Justyn has expertise in product management and technology from his roles at Reflex Media, Inc., where he led product lines in dating, social media, and AI, and at Store Celebrations as the CTO. He has a history of championing disruptive technologies, from his work at The Parker Method, a novel weight loss program, to his tenure at Real Agent Guard, LLC, where he introduced safety technology innovations. Justyn's academic background includes a Masters in Computer Information Systems and a degree in Business Management from the University of Phoenix. In addition to his tech endeavors, Justyn has showcased his strategic acumen in roles at MGM Resorts International, The Venetian Resort Las Vegas, and Central Research. His dedication to innovation and technology-driven solutions has made him a prominent figure in the tech landscape.

Four Investors Explain Why AI Ethics Cannot Be an Afterthought

Deep Nishar

Biography

Deep Nishar is a seasoned technology executive and investor, currently serving as the Managing Director at General Catalyst. Here, he champions positive, enduring change, partnering with founders from seed to growth stages to build institutions that last. Deep's investments span a range of innovative companies, including Adept, Eikon, and Maze Therapeutics. Prior to this, he was a Senior Managing Partner at SoftBank Investment Advisers, where he was responsible for global investments in software, frontier tech, and healthcare, leading to numerous successful IPOs and M&As. Deep's influence in the tech world is also evident from his transformative role at LinkedIn, where as the Senior Vice President of Products & User Experience, he grew the platform's user base from 32 million to 347 million and significantly increased its revenue. Before LinkedIn, Deep made significant contributions at Google, leading product initiatives for the Asia-Pacific region, launching global mobile initiatives, and earning the Google Founders Award for his innovations. Deep co-taught a course on Product Management Fundamentals at Stanford University and has been recognized for his contributions to the tech industry in various publications. He holds an MBA from Harvard Business School, where he graduated as a Baker Scholar, an M.S.E.E. from the University of Illinois Urbana-Champaign, and a B. Tech (Honors) from the Indian Institute of Technology, Kharagpur.

Four Investors Explain Why AI Ethics Cannot Be an Afterthought

Henri Pierre-Jacques

Biography

Henri Pierre-Jacques is Co-Founder and Managing Partner of Harlem Capital, a venture capital firm changing the face of entrepreneurship. He has led 25+ investments and sat on 10+ company boards. Henri formerly sat on the Amazon Black Business Accelerator Board and Management Leadership for Tomorrow NYC Board. Previously, he was a Private Equity Investor at ICV Partners and an Investment Banker in the Real Estate Group at Bank of America. Henri received the Forbes 30 Under 30, Inc 30 Under 30, EBONY Power 100, The Root 100, Business Insider Rising Star, Crains New York Rising Star and HBCUvc 31 Under 31 awards. He has also been featured in WSJ, CNBC, Bloomberg, TechCrunch, Crunchbase, Inc., Black Enterprise and PitchBook. Henri holds an MBA from Harvard Business School and Bachelor of Arts in Economics and minors in Political Science and Business Institutions from Northwestern University. He is Haitian American and originally from Detroit, MI.

Four Investors Explain Why AI Ethics Cannot Be an Afterthought

By General Catalyst, Lightship, Harlem, Angel

Billions of dollars are flooding into AI. Yet, AI models are already being affected by prejudice, as evidenced by mortgage discrimination toward Black prospective homeowners.
It's reasonable to ask what role ethics plays in the building of this technology and, perhaps more importantly, where investors fit in as they rush to fund it.

A founder recently told TechCrunch+ that it's hard to think about ethics when innovation is so rapid: People build systems, then break them, and then edit. So some onus lies on investors to make sure these new technologies are being built by founders with ethics in mind.

To see whether that's happening, TechCrunch+ spoke with four active investors in the space about how they think about ethics in AI and how founders can be encouraged to think more about biases and doing the right thing.

Some investors said they tackle this by doing due diligence on a founder's ethics to help determine whether they'll continue to make decisions the firm can support.

"Founder empathy is a huge green flag for us," said Alexis Alston, principal at Lightship Capital. "Such people understand that while we are looking for market returns, we are also looking for our investments to not cause a negative impact on the globe."

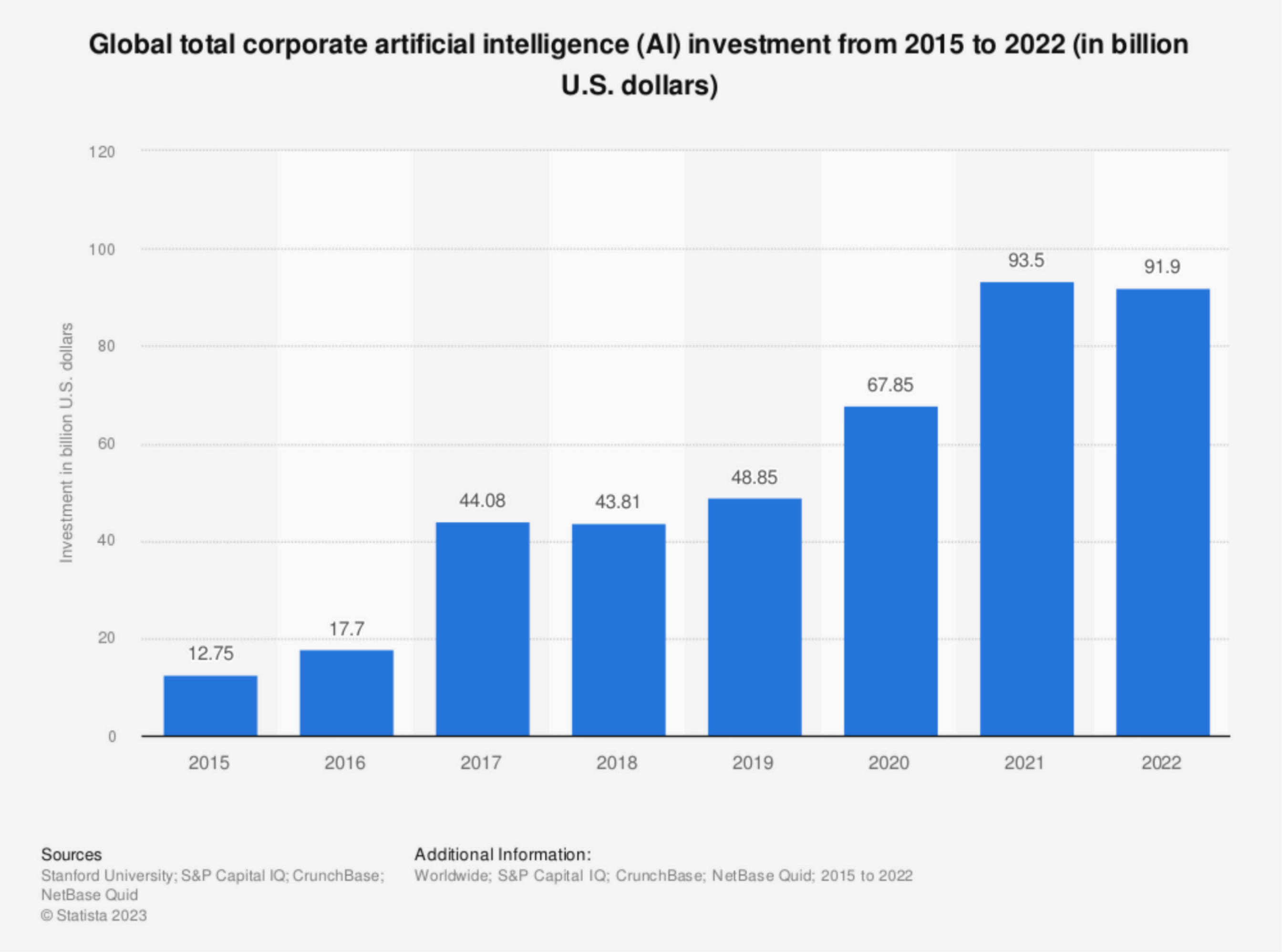
Other investors think that asking hard questions can help separate the wheat from the chaff.

"Any technology brings with it unintended consequences, be it bias, reduced human agency, breaches of privacy or something else," said Deep Nishar, managing director at General Catalyst. "Our investment process centers around identifying such unintended consequences, discussing them with founding teams and assessing whether safeguards are or will be in place to mitigate them."

Government policies are also taking aim at AI: The EU has passed machine learning laws, and the U.S. has introduced plans for an AI task force to start looking at the risks associated with AI. That's in addition to the AI Bill of Rights introduced last year. With many top VC firms injecting money into AI efforts in China, it's important to ask how global ethics within AI can be enforced across borders as well.

We spoke with:

Alexis Alston, principal, Lightship Capital
Justyn Hornor, angel investor and serial founder
Deep Nishar, managing director, General Catalyst
Henri Pierre-Jacques, co-founder and managing partner, Harlem Capital



Investing in AI has increased drastically over the the last decade

Alexis Alston, principal, Lightship Capital

When investing in an AI company, how much due diligence do you do on how its AI model purports or handles bias?
For us, it's important to understand exactly what data the model takes in, where the data comes from and how they're cleaning it. We do quite a bit of technical diligence with our AI-focused GP to make sure that our models can be trained to mitigate or eliminate bias.

We all remember not being able to have faucets turn on automatically to wash our darker hands, and the times when Google image search "accidentally" equated Black skin with primates. I'll do everything in my power to make sure we don't end up with models like that in our portfolio.

How would the U.S. passing machine learning laws similar to the EU's affect the pace of innovation the country sees in this sector?
Given the lack of technical knowledge and sophistication in our government, I have very little faith in the U.S.' ability to pass actionable and accurate legislation around machine learning. We have such a long tail when it comes to timely legislation and for technical experts to be a part of task forces to inform our legislators.

I actually don't see legislation making any major changes in the pace of the development of ML, given how our laws are usually structured. Similarly to the race to the bottom for legislation around designer drugs in the U.S. a decade ago, the legislation never could keep up.

How could and should ethics be defined and enforced globally and across cultures?
We have a deep responsibility to ensure that our investments have no negative implications for national security or contribute to any sort of hyper controlled police state. We have turned down plenty of investments that contribute to the prison industrial complex, threaten national security, or otherwise target marginalized groups of people in a way that elicits harm.

Each firm and each nation has to have its own standard for ethics development, and I don't think there is a blanket framework for AI ethics that would work for all.

Training AI models against discrimination will require expertise beyond engineering. What roles will sociologists, historians, philosophers and other humanities professions play in the future of AI?

I think that sociologists, psychologists and philosophers will play a very large role in these conversations, as they have a deeper understanding of the larger societal implications of legislation and changes in innovation on a global scale than an investor would.

Which sectors of AI seem to be ahead of the rest when it comes to adopting ethical oversight? Which could use more of it?
Facial recognition is likely the furthest along, given its tenure as an established area in ML that has had strong leadership in communities of color and was led by engineers of color for years. Many of these teams were at the forefront of [studying] the implications of AI in immigration, policing and other policy-driven initiatives.

Every other aspect of AI could use deeper ethical oversight, especially the use of AI in predictive policing, drone deployment and most defense uses. All of these are areas I would not be comfortable driving AI innovation in.

What are the red and green flags you look for when investing in an AI product with regard to ethical considerations?
Founder empathy is a huge green flag for us, as such people understand that while we are looking for market returns, we are also looking for our investments to not cause a negative impact on the globe. Diversity of team and thought, particularly in product and engineering, is also crucial, as these teams have to have a keen eye for factors that can negatively impact algorithmic development.

Red flags for us are homogeneous teams or a general lack of forethought or accountability around the larger implications of AI, machine learning and computer vision.

What is investors' responsibility for ensuring that ethics stays at the forefront of the conversation surrounding innovation within AI?
I think we all have a deep responsibility here — funders, founders, operators, policymakers and thought leaders in sociology — to ensure that AI and ethics go hand in hand. We've been having these conversations for years now, and I'm glad that ChatGPT is bringing it back to the forefront of people's minds for us to collectively work toward a more equitable and safe future.

Justyn Hornor, angel investor and serial founder

When investing in an AI company, how much due diligence do you do on how its AI model purports or handles bias?

I look for two key elements:

- Are the key risks for bias clearly understood and measured?
- Does the system include human-in-the-loop capabilities for constant learning?

Bias is going to be very specific to the model and its use cases, and the risks are going to be highly dependent upon the industry. For medical-based AI products, for example, these risks would be examined with a great deal of detail. But a manufacturing system where AI is used to determine the quality of a bolt will not need the same level of scrutiny.

How would the U.S. passing machine learning laws similar to the EU's affect the pace of innovation the country sees in this sector?
The pace may be slowed in the short term, but markets will adapt quickly with standards and systems that will be commoditized in short order. I believe these laws in the EU are well designed and are being thoughtfully implemented.

We should anticipate similar laws in the near future and begin self-policing our respective industries and systems to get ahead of these regulatory changes.

How could and should ethics be defined and enforced globally and across cultures?
These are two huge questions. Global ethics are typically very, very high level. That doesn't mean they're not valuable, but abstracting an ethical framework for AI products at such a high level may lead to an inability to enforce those standards. Many of the bigger challenges with regard to China can be addressed through non-AI ethics frameworks. For example, consumer privacy and theft of intellectual property are clearly defined in most modern markets.

Enforcing standards with regard to China will likely come from major trade pacts. The [Trans-Pacific Partnership] was a particularly powerful approach until it got pulled into culture war nonsense here in the U.S. Outside of major multilateral trade agreements, there are few means of enforcing any kind of international standards with China outside of saber-rattling or war.

Training AI models against discrimination will require expertise beyond engineering. What roles will sociologists, historians, philosophers and other humanities professions play in the future of AI?
We will see multidisciplinary teams become the norm with regard to AI training. There are a couple of facets of interest here depending on the type of AI products being built.

For generative AI, these systems will have to find a balance between responding with "correct" information and being prompted by humans with their own biases. Many historical events can be viewed through a number of different viewpoints, for example. It will be challenging to find a sweet spot, especially when humans can exert a significant influence on the outputs of generative AI: text, video, audio, images, etc.

For classification systems, I believe we'll see similar teams that will face challenges in how the various labels influence the outputs of their AI. A common use case that I have run into many times is AI vision products that treat nudity in art in the same way as pornography. That's a form of bias that must be understood within the context of a culture. There are no easy answers.

What are the red and green flags you look for when investing in an AI product with regard to ethical considerations?
I want to see that the team has been mindful and deliberate about defining, measuring and controlling

biases. They may not get it right, but being intentional is very important.

Additionally, I want to understand their source of underlying data used for training. Besides the direct sources, feature engineering is a common means of extrapolating data from primary sources, so I want to understand if and how this process has been applied to any training.

What is investors' responsibility for ensuring that ethics stays at the forefront of the conversation surrounding innovation within AI?
Investors should be asking the hard questions early. If you don't have expertise on the team to understand the systems, hire subject matter experts to help dig into the technology. You may find a lot of red flags — that doesn't mean you shouldn't invest; just make sure the use of funds includes elevating the systems being built.

I also believe that any company with products or services that have significant amounts of AI running systems should have a trust and safety executive at or near the C-suite. There should be oversight of these systems and someone on the ground in the company with access to engineering teams who can also access the C-suite without risks of raising concerns.

Investors should be pushing for these roles and accept that use of funds include onboarding this type of expertise.

Deep Nishar, managing director, General Catalyst

When investing in an AI company, how much due diligence do you do on how its AI model purports or handles bias?
Bias is one of many dimensions within ethical AI (or responsible AI, as we refer to it at GC) that we evaluate in every investment decision we make. The idea of ethical AI cuts across our three primary responsible innovation pillars of inclusive prosperity, sustainable development and good citizenship. We believe this framework is a toolkit for us and our companies: it extends beyond due diligence into scaling companies and scoping second/third acts.
Any technology brings with it unintended consequences, be it bias, reduced human agency, breaches of privacy or something else. Our investment process centers around identifying such unintended consequences, discussing them with founding teams and assessing whether safeguards are or will be in place to mitigate them.

Any technology brings with it unintended consequences, be it bias, reduced human agency, breaches of privacy or something else. Our investment process centers around identifying such unintended consequences, discussing them with founding teams and assessing whether safeguards are or will be in place to mitigate them.

How would the U.S. passing machine learning laws similar to the EU's affect the pace of innovation the country sees in this sector?
Across history, we see policy impacts on innovation occupy a spectrum. We've seen too little thus far to diagnose Capitol Hill's influence on AI's direction of travel.

That said, basic measures of fairness, transparency, privacy and reliability should be instituted with standard protocols and methodologies backing them. We believe that if guidelines are meant to be universal, compliance therein must be universally accessible and understandable. We believe the first step is a modicum of standard transparency that, similar to nutrition labels, will afford users knowledge of what they are consuming.

Training AI models against discrimination will require expertise beyond engineering. What roles will sociologists, historians, philosophers and other humanities professions play in the future of AI?
Both direct and indirect roles for the arts and humanities exist in this AI era, and they fill critical gaps in purely technical reasoning. It is perhaps easiest to envisage them at the inception and terminus of AI workflows: Does an architecture reflect the intentions of the architect (and society)? Do the inputs fed to the architecture holistically represent intentions and population(s)? Are these results aligned with the intentions outlined at inception? Of what consequence may they be to broader populations?

What are the red and green flags you look for when investing in an AI product with regard to ethical considerations?
Every investment memo we write includes a diagnostic on principles of responsible innovation. To this end, we have conversations with founders about this by the time we get to a term sheet. For us, ethical AI is about fostering the right mindsets and mechanisms such that responsible AI emanates from the core of the company. We probe for extant safeguards at the technological and organizational levels, and have discussions with teams about the potential unintended consequences of their products.

It's a red flag if our conversations with teams demarcate the first time these topics have surfaced.

What is investors' responsibility for ensuring that ethics stays at the forefront of the conversation surrounding innovation within AI?
Our fundamental belief is that every stakeholder — technologist or not, from builders to end users — plays a role in ethical AI.

At the end of the day, we vote with our checkbooks and our governance rights. We believe that ethical AI and financial returns are not in competition with one another — quite the opposite, actually.

Responsible and ethical innovation contributes to stronger and more enduring companies, which in turn leads to better investment outcomes. To that end, the way we see it, ethical AI is a natural extension of the fiduciary duties by which we are already bound.

Henri Pierre-Jacques, managing partner, Harlem Capital

When investing in an AI company, how much due diligence do you do on how its AI model purports or handles bias?
Given we invest at the pre-seed and seed stage, most of the AI companies at that point are pre-product or pre-revenue, so it's very early in the tech product development. We are [doing due diligence on] the founder's ethics to determine if they will make decisions we support over many years.

How would the U.S. passing machine learning laws similar to the EU's affect the pace of innovation the country sees in this sector?
Innovation won't be stopped, just altered. Whether it's the EU or China, both have stricter rules, but both are still innovating. The right balance of laws is still unclear.

How could and should ethics be defined and enforced globally and across cultures?
Every country and region will have to make that decision for themselves; no one knows the right solution at this point, as everyone is just figuring it out. In reality, corporations will make decisions ahead of governments in most regions.

Training AI models against discrimination will require expertise beyond engineering. What roles

will sociologists, historians, philosophers and other humanities professions play in the future of AI?
In an ideal world, they would work similarly to a marketing and engineering team, but I don't have a lot of hope that this will be the case, as it wasn't for web3, either.

Which sectors of AI seem to be ahead of the rest when it comes to adopting ethical oversight? Which could use more of it?
AI for [autonomous vehicles] has been a long and slow rollout. They have spent time thinking about insurance, death, regulation, job loss and more. The rollout of generative AI for consumer-facing products like images or chatbots has felt like it's gone really fast.

The stakes seem lower at first, because death by a car accident isn't an option, but there are still some serious implications from the consumer-facing technology that hasn't been fully thought through, in my opinion.

What is investors' responsibility for ensuring that ethics stays at the forefront of the conversation surrounding innovation within AI?
Given the power of this technology shift, I think it's critical. We believe that companies should be making governance decisions even if their governments don't require it.

KEY TAKEAWAYS

• There are increasing concerns about ethics in AI Investments as AI models are being affected by biases. Rapid innovation makes it challenging to consider ethics, putting responsibility on investors to ensure ethical considerations. Government policies, like the EU's machine learning laws and the U.S.'s plans for an AI task force, are targeting AI ethics.

- **Investor Perspectives on Ethical AI:**
  - Alexis Alston (Lightship Capital):
    - Emphasizes understanding the data an AI model uses and ensuring it's free from bias.
    - Believes U.S. legislation might not significantly impact the pace of AI development.
    - Advocates for a diverse team and founder empathy as green flags.
  - Justyn Hornor (Angel Investor):
    - Stresses the importance of understanding and controlling biases in AI models.
    - Believes U.S. should anticipate and adapt to machine learning laws similar to the EU's.
    - Advocates for multidisciplinary teams in AI training.
  - Deep Nishar (General Catalyst):
    - Focuses on identifying unintended consequences of AI and ensuring safeguards are in place.
    - Believes in fostering responsible AI from the core of a company.
  - Henri Pierre-Jacques (Harlem Capital):
    - Emphasizes the importance of founder ethics in early-stage AI companies.
    - Believes innovation will continue despite regulations, but its nature might change.

- **Challenges and Implications of Ethical AI:**
  - Training AI models against discrimination will require expertise from humanities professions like sociologists, historians, and philosophers.
  - Different sectors of AI have varying levels of ethical oversight. For instance, AI for autonomous vehicles has been more deliberate in its rollout compared to consumer-facing AI products.
  - Ethical considerations in AI are not just about avoiding biases but also about understanding the broader societal implications:

- **Role of Investors in Promoting Ethical AI:**
  - Investors have a responsibility to ensure that ethics remain a primary consideration in AI innovation.
  - They should be proactive in asking hard questions and understanding the systems they're investing in.
  - Ethical AI is seen as contributing to stronger, more enduring companies, aligning with better investment outcomes.

The Next Big Opportunity for Venture Capital Is Not Based on AI, but on Trust

Tracy Barba

Biography

Tracy Barba is the founder of Responsible VC, a nonprofit advisory and consultancy advancing responsible policies and practices in venture capital. She was most recently head of ESG and stakeholder engagement at 500 Global. Before that, she led capital formation for Bamboo Capital Partners, a private equity fund investing in financial services, renewable energy, and healthcare sectors across more than 30 emerging and frontier market countries. Previously, she led public affairs at Social Finance US, the firm behind the Social Impact Bond. She was the West Coast executive director of Golden Seeds, a 275-member investor network investing in women-led companies. She started her career with venture funds in Silicon Valley, including VantagePoint Venture Partners and InterWest Partners.

The Next Big Opportunity for Venture Capital Is Not Based on AI, but on Trust

By Tracy Barba

Venture capital (VC) has long been the crucible of innovation, fueling startups that challenge the status quo and revolutionize industries. These investments have not just catalyzed business productivity but have played a pivotal role in sculpting modern-day economic landscapes. One such frontier has been the realm of Generative AI, which in 2023 alone, has seen a meteoric rise in VC investments, recording an almost five-fold increase from last year's figures.

But as the prospects of AI beckon, an underlying ethical conundrum emerges. The venture industry, once celebrated for its 'move fast and break things' mantra, now stands at a crossroads, weighing the unforeseen repercussions of AI.

Notably, concerns range from the ethical opacity of "black box" AI outputs to the looming shadows of Big Tech dominance, casting a shadow over the industry's vibrancy. A prevalent sentiment resonating in the recent edition of Equation, The Tech Ethics Quarterly posits a glaring issue – trust, or rather the lack of it, is becoming a formidable barrier to AI's pervasive adoption.

While AI's promises are vast, encompassing sectors from healthcare to finance, the challenges are equally pronounced. Industry heavyweights like Sequoia Capital and figures like Marc Andreessen have been vocal about their AI enthusiasm. Coatue has similarly projected a bold vision of AI, touting its capabilities to amplify human productivity and provide new avenues to explore the human experience.

Nevertheless, this optimism is countered by sobering realities. Generative AI, for instance, has exposed the tech community to various issues, from misinformation and biased data to the lack of transparency in AI algorithms. In this labyrinth of technological prowess and pitfalls, how can venture capitalists strike the right balance?

Policy and regulation will undoubtedly play a decisive role. The European Union, in a pioneering move, has unveiled the AI Act, reminiscent of the seismic influence GDPR once had on data privacy. As nations like China, India, and Australia prepare to embrace AI's maturation, they, too, are contemplating regulatory interventions.

Interestingly, despite a drop in the founding of new AI enablement firms, emerging ventures are methodical and grounded in real needs rather than mere hype. This underscores a realization that for AI to be successful, it has to be trustworthy.

The AI Ethics Maturity Continuum, a culmination of collective insights from Ethical Intelligence, BGV, and EAIGG, emerges as a pivotal tool for both startups and venture capitalists (VCs). Rooted in industry best practices, academic contributions, and AI policy guidelines, this framework proves invaluable for:

• Investors who wish to use the assessment to streamline their due diligence (DD) processes. The Continuum not only offers a snapshot of an AI venture's ethical health, but also serves as a practical instrument in the evaluation process.

- VCs aiming to liaise with their boards to embed best practices at the executive and managerial tiers of their invested companies. The Continuum becomes a roadmap, anchored in KPIs, guiding them in fortifying ethical principles within their portfolio's leadership and operations.

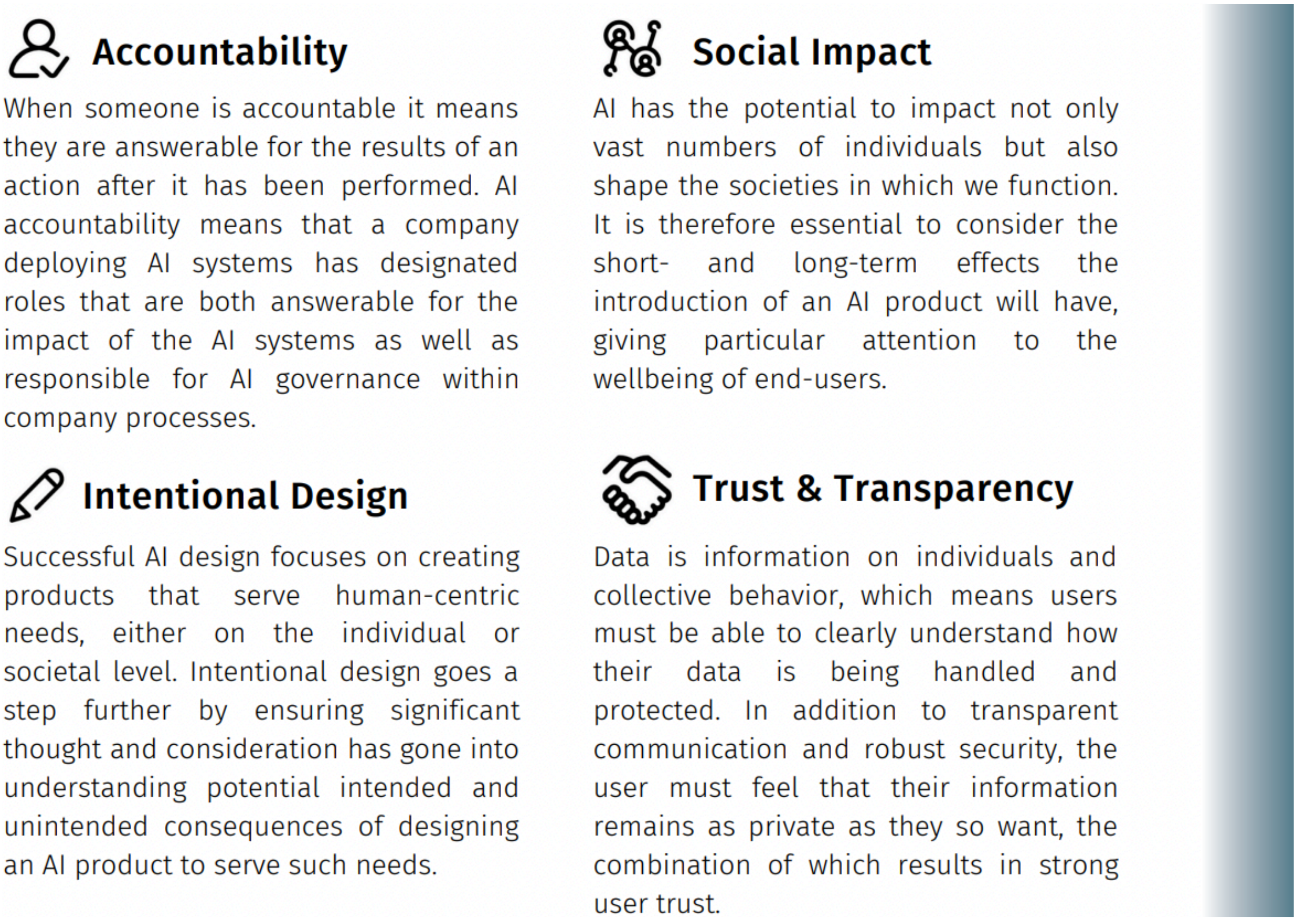
- Those who are keen on monitoring their portfolio's progression, based on pivotal metrics, feedback, and other essential indicators, ensuring that ethics remains a cornerstone of business growth and innovation.

In this day and age, successful operationalization of ethics in AI-driven startups has become the competitive edge. By striving to design and develop AI with ethics at the core, startups can maximize their impact and continue to grow in a market environment where values are becoming increasingly important to success.

To aid in this process, the Ethics Maturity Continuum has been designed to quickly assess a company's level of ethics maturity and identify areas for improvement. It prioritizes agility and action, enabling users to build concrete strategies for sustainable AI systems and track development overtime. Most importantly, it empowers startups to embed ethics from the very beginning, resulting in stronger products, happier customers and more favorable exits.

Delving deeper, the Ethics Maturity Continuum assessment is structured around five pivotal questions that orbit key ethical pillars:

The Continuum Framework



Accountability

When someone is accountable it means they are answerable for the results of an action after it has been performed. AI accountability means that a company deploying AI systems has designated roles that are both answerable for the impact of the AI systems as well as responsible for AI governance within company processes.

Social Impact

AI has the potential to impact not only vast numbers of individuals but also shape the societies in which we function. It is therefore essential to consider the short- and long-term effects the introduction of an AI product will have, giving particular attention to the wellbeing of end-users.

Intentional Design

Successful AI design focuses on creating products that serve human-centric needs, either on the individual or societal level. Intentional design goes a step further by ensuring significant thought and consideration has gone into understanding potential intended and unintended consequences of designing an AI product to serve such needs.

Trust & Transparency

Data is information on individuals and collective behavior, which means users must be able to clearly understand how their data is being handled and protected. In addition to transparent communication and robust security, the user must feel that their information remains as private as they so want, the combination of which results in strong user trust.

⚖ Fairness

Unwanted bias occurs when system based decisions are made using individual traits that should not otherwise correlate to the outcome (i.e. gender being used as a deciding factor for job applicants). Fairness seeks to minimize instances of this unwanted bias and instead promote inclusive representation in AI development.

The Assessment is designed across two cohorts: 1) for Early Stage companies, up to Series B; and 2) For Late Stage, Series C and beyond. Each of the aforementioned categories are then evaluated in grounded and concrete terms, measuring the level of maturity, and identifying clear actions for improvement, with KPIs and a decisionmaker on the team.

Through its deployment, EAIGG conducted a benchmark analysis involving 15 venture-backed startups in Silicon Valley. The findings amassed were illuminating, furnishing invaluable insights for future ethical considerations and strategies.

The cohort of companies showed moderate emphasis on trust and transparency, and fairness, with average scores of 13.66 out of 20 in both categories. There's a slightly higher commitment to AI governance with an average score of 22.66 out of 30 in accountability. However, intentional design and social impact have room for improvement, scoring 9 out of 15 and 11.33 out of 15 respectively. This demonstrated that startups may need to invest more heavily in having a product manager or leader to review unintended consequences of their product, and designate a clear mission statement aligned with responsible AI.

For venture capitalists, the writing on the wall is clear. Their role transcends merely hunting the next billion-dollar unicorn; it's about stewarding an era where technology and morality converge.

Enter the Responsible AI VC Council, an initiative of the Lucas Institute for Venture Ethics at the Markkula Center for Applied Ethics. The council, supported by the Lucas Brothers Foundation, seeks to foster dialogue, disseminate best practices, and provide the venture community with insights into AI legislation, tools, and risk management frameworks. Donald Lucas, often hailed as the godfather of venture capital, symbolizes the ethos of this endeavor. The goal? To marry the pioneering spirit of venture capitalism with the rigorous ethical standards the age of AI necessitates.

As the AI tapestry unfolds, it's evident that trust will be its cornerstone. For venture capitalists, the call to action is twofold: to invest and innovate, but also to safeguard the ethical underpinnings of a technology poised to redefine our world. The intersection of innovation and integrity is where the future of venture capital lies.

To learn more about Responsible AI VC Council (insert link to website). To take the Ethics Maturity Continuum Assessment, click here.

KEY TAKEAWAYS

- **Shift in Venture Capital Focus:** While Generative AI has seen a significant rise in VC investments, there's a growing ethical concern surrounding AI. The venture industry is now grappling with the challenges of "black box" AI outputs, Big Tech dominance, and a lack of trust, which is becoming a major barrier to AI's widespread adoption.

- **Balancing AI's Promise and Pitfalls:** Despite the vast potential of AI across sectors, there are significant challenges, including misinformation, biased data, and a lack of transparency in AI algorithms. The article emphasizes the need for venture capitalists to find a balance between technological advancement and ethical considerations.

- **Regulatory Interventions:** Policy and regulation will play a crucial role in shaping the AI landscape. The European Union has introduced the AI Act, and other nations like China, India, and Australia are also considering regulatory measures.

- **AI Ethics Maturity Continuum:** This tool, developed from insights from Ethical Intelligence, BGV, and EAIGG, serves as a guide for startups and VCs. It helps in assessing an AI venture's ethical health, embedding best practices at executive and managerial levels, and ensuring ethics remains central to business growth and innovation. The Continuum focuses on five key ethical pillars: Accountability, Intentional Design, Fairness, Social Impact, and Trust & Transparency.

- **Responsible AI VC Council:** The Lucas Institute for Venture Ethics at the Markkula Center for Applied Ethics has launched the Responsible AI VC Council. Supported by the Lucas Brothers Foundation, the council aims to promote dialogue, share best practices, and provide insights into AI legislation, tools, and risk management frameworks. The overarching goal is to combine the innovative spirit of venture capitalism with the ethical standards required in the AI era.

In essence, the future of venture capital is seen at the intersection of innovation and integrity, with trust being the foundational element.

Why AI Will Save the World

Marc Andreessen

Biography

Marc Andreessen is a Cofounder and General Partner at the venture capital firm Andreessen Horowitz. He is an innovator and creator, one of the few to pioneer a software category used by more than a billion people and one of the few to establish multiple billion-dollar companies. Marc co-created the highly influential Mosaic internet browser and co-founded Netscape, which later sold to AOL for $4.2 billion. He also co-founded Loudcloud, which as Opsware, sold to Hewlett-Packard for $1.6 billion. He later served on the board of Hewlett-Packard from 2008 to 2018. Marc holds a BS in Computer Science from the University of Illinois at Urbana-Champaign. Marc serves on the board of the following Andreessen Horowitz portfolio companies: Applied Intuition, Carta, Coinbase, Dialpad, Flow, Golden, Honor, OpenGov, and Samsara. He is also on the board of Meta.

Why AI Will Save the World

Marc Andreessen

Source: https://a16z.com/ai-will-save-the-world/

The era of Artificial Intelligence is here, and boy are people freaking out.

Fortunately, I am here to bring the good news: AI will not destroy the world, and in fact may save it.

First, a short description of what AI is: The application of mathematics and software code to teach computers how to understand, synthesize, and generate knowledge in ways similar to how people do it. AI is a computer program like any other – it runs, takes input, processes, and generates output. AI's output is useful across a wide range of fields, ranging from coding to medicine to law to the creative arts. It is owned by people and controlled by people, like any other technology.

A shorter description of what AI isn't: Killer software and robots that will spring to life and decide to murder the human race or otherwise ruin everything, like you see in the movies.

An even shorter description of what AI could be: A way to make everything we care about better.

Why AI Can Make Everything We Care About Better

The most validated core conclusion of social science across many decades and thousands of studies is that human intelligence makes a very broad range of life outcomes better. Smarter people have better outcomes in almost every domain of activity: academic achievement, job performance, occupational status, income, creativity, physical health, longevity, learning new skills, managing complex tasks, leadership, entrepreneurial success, conflict resolution, reading comprehension, financial decision making, understanding others' perspectives, creative arts, parenting outcomes, and life satisfaction.

Further, human intelligence is the lever that we have used for millennia to create the world we live in today: science, technology, math, physics, chemistry, medicine, energy, construction, transportation, communication, art, music, culture, philosophy, ethics, morality. Without the application of intelligence on all these domains, we would all still be living in mud huts, scratching out a meager existence of subsistence farming. Instead we have used our intelligence to raise our standard of living on the order of 10,000X over the last 4,000 years.

What AI offers us is the opportunity to profoundly augment human intelligence to make all of these outcomes of intelligence – and many others, from the creation of new medicines to ways to solve climate change to technologies to reach the stars – much, much better from here.

AI augmentation of human intelligence has already started – AI is already around us in the form of computer control systems of many kinds, is now rapidly escalating with AI Large Language Models like ChatGPT, and will accelerate very quickly from here – *if we let it*.

In our new era of AI:

- Every child will have an AI tutor that is infinitely patient, infinitely compassionate, infinitely knowledgeable, infinitely helpful. The AI tutor will be by each child's side every step of their development, helping them maximize their potential with the machine version of infinite love.
- Every person will have an AI assistant/coach/mentor/trainer/advisor/therapist that is infinitely patient, infinitely compassionate, infinitely knowledgeable, and infinitely helpful. The AI assistant will be present through all of life's opportunities and challenges, maximizing every person's outcomes.
- Every scientist will have an AI assistant/collaborator/partner that will greatly expand their scope of scientific research and achievement. Every artist, every engineer, every businessperson, every doctor, every caregiver will have the same in their worlds.
- Every leader of people – CEO, government official, nonprofit president, athletic coach, teacher – will have the same. The magnification effects of better decisions by leaders across the people they lead are enormous, so this intelligence augmentation may be the most important of all.
- Productivity growth throughout the economy will accelerate dramatically, driving economic growth, creation of new industries, creation of new jobs, and wage growth, and resulting in a new era of heightened material prosperity across the planet.
- Scientific breakthroughs and new technologies and medicines will dramatically expand, as AI helps us further decode the laws of nature and harvest them for our benefit.
- The creative arts will enter a golden age, as AI-augmented artists, musicians, writers, and filmmakers gain the ability to realize their visions far faster and at greater scale than ever before.
- I even think AI is going to improve warfare, when it has to happen, by reducing wartime death rates dramatically. Every war is characterized by terrible decisions made under intense pressure and with sharply limited information by very limited human leaders. Now, military commanders and political leaders will have AI advisors that will help them make much better strategic and tactical decisions, minimizing risk, error, and unnecessary bloodshed.
- In short, anything that people do with their natural intelligence today can be done much better with AI, and we will be able to take on new challenges that have been impossible to tackle without AI, from curing all diseases to achieving interstellar travel.
- And this isn't just about intelligence! Perhaps the most underestimated quality of AI is how humanizing it can be. AI art gives people who otherwise lack technical skills the freedom to

> create and share their artistic ideas. Talking to an empathetic AI friend really does improve their ability to handle adversity. And AI medical chatbots are already more empathetic than their human counterparts. Rather than making the world harsher and more mechanistic, infinitely patient and sympathetic AI will make the world warmer and nicer.

The stakes here are high. The opportunities are profound. AI is quite possibly the most important – and best – thing our civilization has ever created, certainly on par with electricity and microchips, and probably beyond those.

The development and proliferation of AI – far from a risk that we should fear – is a moral obligation that we have to ourselves, to our children, and to our future.

We should be living in a much better world with AI, and now we can.

So Why The Panic?

In contrast to this positive view, the public conversation about AI is presently shot through with hysterical fear and paranoia.

We hear claims that AI will variously kill us all, ruin our society, take all our jobs, cause crippling inequality, and enable bad people to do awful things.

What explains this divergence in potential outcomes from near utopia to horrifying dystopia?

Historically, every new technology that matters, from electric lighting to automobiles to radio to the Internet, has sparked a moral panic – a social contagion that convinces people the new technology is going to destroy the world, or society, or both. The fine folks at Pessimists Archive have documented these technology-driven moral panics over the decades; their history makes the pattern vividly clear. It turns out this present panic is not even the first for AI.

Now, it is certainly the case that many new technologies have led to bad outcomes – often the same technologies that have been otherwise enormously beneficial to our welfare. So it's not that the mere existence of a moral panic means there is nothing to be concerned about.

But a moral panic is by its very nature irrational – it takes what may be a legitimate concern and inflates it into a level of hysteria that ironically makes it harder to confront actually serious concerns.

And wow do we have a full-blown moral panic about AI right now.

This moral panic is already being used as a motivating force by a variety of actors to demand policy action – new AI restrictions, regulations, and laws. These actors, who are making extremely dramatic public statements about the dangers of AI – feeding on and further inflaming moral panic – all present themselves as selfless champions of the public good.

But are they?
And are they right or wrong?

The Baptists And Bootleggers Of AI

Economists have observed a longstanding pattern in reform movements of this kind. The actors within movements like these fall into two categories – "Baptists" and "Bootleggers" – drawing on the historical example of the prohibition of alcohol in the United States in the 1920's:

Baptists
"Baptists" are the true believer social reformers who legitimately feel – deeply and emotionally, if not rationally – that new restrictions, regulations, and laws are required to prevent societal disaster. For alcohol prohibition, these actors were often literally devout Christians who felt that alcohol was destroying the moral fabric of society. For AI risk, these actors are true believers that AI presents one or another existential risks – strap them to a polygraph, they really mean it.

Bootleggers
"Bootleggers" are the self-interested opportunists who stand to financially profit by the imposition of new restrictions, regulations, and laws that insulate them from competitors. For alcohol prohibition, these were the literal bootleggers who made a fortune selling illicit alcohol to Americans when legitimate alcohol sales were banned. For AI risk, these are CEOs who stand to make more money if regulatory barriers are erected that form a cartel of government-blessed AI vendors protected from new startup and open source competition – the software version of "too big to fail" banks.

A cynic would suggest that some of the apparent Baptists are also Bootleggers – specifically the ones paid to attack AI by their universities, think tanks, activist groups, and media outlets. If you are paid a salary or receive grants to foster AI panic…you are probably a Bootlegger.

The problem with the Bootleggers is that they win. The Baptists are naive ideologues, the Bootleggers are cynical operators, and so the result of reform movements like these is often that the Bootleggers get what they want – regulatory capture, insulation from competition, the formation of a cartel – and the Baptists are left wondering where their drive for social improvement went so wrong.

We just lived through a stunning example of this – banking reform after the 2008 global financial crisis. The Baptists told us that we needed new laws and regulations to break up the "too big to fail" banks to prevent such a crisis from ever happening again. So Congress passed the Dodd-Frank Act of 2010, which was marketed as satisfying the Baptists' goal, but in reality was coopted by the Bootleggers – the big banks. The result is that the same banks that were "too big to fail" in 2008 are much, much larger now.

So in practice, even when the Baptists are genuine – and even when the Baptists are right – they are used as cover by manipulative and venal Bootleggers to benefit themselves.

And this is what is happening in the drive for AI regulation right now.

However, it isn't sufficient to simply identify the actors and impugn their motives. We should consider the arguments of both the Baptists and the Bootleggers on their merits.

AI Risk #1: Will AI Kill Us All?

The first and original AI doomer risk is that AI will decide to literally kill humanity.

The fear that technology of our own creation will rise up and destroy us is deeply coded into our culture. The Greeks expressed this fear in the Prometheus Myth – Prometheus brought the destructive power of fire, and more generally technology ("techne"), to man, for which Prometheus was condemned to perpetual torture by the gods. Later, Mary Shelley gave us moderns our own version of this myth in her novel Frankenstein, or, The Modern Prometheus, in which we develop the technology for eternal life, which then rises up and seeks to destroy us. And of course, no AI panic newspaper story is complete without a still image of a gleaming red-eyed killer robot from James Cameron's Terminator films.

The presumed evolutionary purpose of this mythology is to motivate us to seriously consider potential risks of new technologies – fire, after all, can indeed be used to burn down entire cities. But just as fire was also the foundation of modern civilization as used to keep us warm and safe in a cold and hostile world, this mythology ignores the far greater upside of most – all? – new technologies, and in practice inflames destructive emotion rather than reasoned analysis. Just because premodern man freaked out like this doesn't mean we have to; we can apply rationality instead.

My view is that the idea that AI will decide to literally kill humanity is a profound category error. AI is not a living being that has been primed by billions of years of evolution to participate in the battle for the survival of the fittest, as animals are, and as we are. It is math – code – computers, built by people, owned by people, used by people, controlled by people. The idea that it will at some point develop a mind of its own and decide that it has motivations that lead it to try to kill us is a superstitious handwave.

In short, AI doesn't want, it doesn't have goals, it doesn't want to kill you, because it's not alive. And AI is a machine – is not going to come alive any more than your toaster will.

Now, obviously, there are true believers in killer AI – Baptists – who are gaining a suddenly stratospheric amount of media coverage for their terrifying warnings, some of whom claim to have been studying the topic for decades and say they are now scared out of their minds by what they have learned. Some of these true believers are even actual innovators of the technology. These actors are arguing for a variety of bizarre and extreme restrictions on AI ranging from a ban on AI development, all the way up to military airstrikes on datacenters and nuclear war. They argue that because people like me cannot rule out future catastrophic consequences of AI, that we must assume a precautionary stance that may require large amounts of physical violence and death in order to prevent potential existential risk.

My response is that their position is non-scientific – What is the testable hypothesis? What would falsify the hypothesis? How do we know when we are getting into a danger zone? These questions go mainly unanswered apart from "You can't prove it won't happen!" In fact, these Baptists' position is so non-scientific and so extreme – a conspiracy theory about math and code – and is already calling for physical violence, that I will do something I would normally not do and question their motives as well.

Specifically, I think three things are going on:

First, recall that John Von Neumann responded to Robert Oppenheimer's famous hand-wringing about his role creating nuclear weapons – which helped end World War II and prevent World War III – with, "Some people confess guilt to claim credit for the sin." What is the most dramatic way one can claim credit for the importance of one's work without sounding overtly boastful? This explains the mismatch between the words and actions of the Baptists who are actually building and funding AI – watch their actions, not their words. (Truman was harsher after meeting with Oppenheimer: "Don't let that crybaby in here again.")

Second, some of the Baptists are actually Bootleggers. There is a whole profession of "AI safety expert", "AI ethicist", "AI risk researcher". They are paid to be doomers, and their statements should be processed appropriately.

Third, California is justifiably famous for our many thousands of cults, from EST to the Peoples Temple, from Heaven's Gate to the Manson Family. Many, although not all, of these cults are harmless, and maybe even serve a purpose for alienated people who find homes in them. But some are very dangerous indeed, and cults have a notoriously hard time straddling the line that ultimately leads to violence and death.

And the reality, which is obvious to everyone in the Bay Area but probably not outside of it, is that "AI risk" has developed into a cult, which has suddenly emerged into the daylight of global press attention and the public conversation. This cult has pulled in not just fringe characters, but also some actual industry experts and a not small number of wealthy donors – including, until recently, Sam Bankman-Fried. And it's developed a full panoply of cult behaviors and beliefs.

This cult is why there are a set of AI risk doomers who sound so extreme – it's not that they actually have secret knowledge that make their extremism logical, it's that they've whipped themselves into a frenzy and really are…extremely extreme.

It turns out that this type of cult isn't new – there is a longstanding Western tradition of millenarianism, which generates apocalypse cults. The AI risk cult has all the hallmarks of a millenarian apocalypse cult. From Wikipedia, with additions by me:
"Millenarianism is the belief by a group or movement [AI risk doomers] in a coming fundamental transformation of society [the arrival of AI], after which all things will be changed [AI utopia, dystopia, and/or end of the world]. Only dramatic events [AI bans, airstrikes on datacenters, nuclear strikes on unregulated AI] are seen as able to change the world [prevent AI] and the change is anticipated to be brought about, or survived, by a group of the devout and dedicated. In most millenarian scenarios, the disaster or battle to come [AI apocalypse, or its prevention] will be followed by a new, purified world [AI bans] in which the believers will be rewarded [or at least acknowledged to have been correct all along]."

This apocalypse cult pattern is so obvious that I am surprised more people don't see it.

Don't get me wrong, cults are fun to hear about, their written material is often creative and fascinating, and their members are engaging at dinner parties and on TV. But their extreme beliefs should not determine the future of laws and society – obviously not.

AI Risk #2: Will AI Ruin Our Society?

The second widely mooted AI risk is that AI will ruin our society, by generating outputs that will be so "harmful", to use the nomenclature of this kind of doomer, as to cause profound damage to humanity, even if we're not literally killed.

Short version: If the murder robots don't get us, the hate speech and misinformation will.

This is a relatively recent doomer concern that branched off from and somewhat took over the "AI risk" movement that I described above. In fact, the terminology of AI risk recently changed from "AI safety" – the term used by people who are worried that AI would literally kill us – to "AI alignment" – the term used by people who are worried about societal "harms". The original AI safety people are frustrated by this shift, although they don't know how to put it back in the box – they now advocate that the actual AI risk topic be renamed "AI notkilleveryoneism", which has not yet been widely adopted but is at least clear.

The tipoff to the nature of the AI societal risk claim is its own term, "AI alignment". Alignment with what? Human values. Whose human values? Ah, that's where things get tricky.

As it happens, I have had a front row seat to an analogous situation – the social media "trust and safety" wars. As is now obvious, social media services have been under massive pressure from governments and activists to ban, restrict, censor, and otherwise suppress a wide range of content for many years. And the same concerns of "hate speech" (and its mathematical counterpart, "algorithmic bias") and "misinformation" are being directly transferred from the social media context to the new frontier of "AI alignment".

My big learnings from the social media wars are:

On the one hand, there is no absolutist free speech position. First, every country, including the United States, makes at least some content illegal. Second, there are certain kinds of content, like child pornography and incitements to real world violence, that are nearly universally agreed to be off limits – legal or not – by virtually every society. So any technological platform that facilitates or generates content – speech – is going to have some restrictions.

On the other hand, the slippery slope is not a fallacy, it's an inevitability. Once a framework for restricting even egregiously terrible content is in place – for example, for hate speech, a specific hurtful word, or for misinformation, obviously false claims like "the Pope is dead" – a shockingly broad range of government agencies and activist pressure groups and nongovernmental entities will kick into gear and demand ever greater levels of censorship and suppression of whatever speech they view as threatening to society and/or their own personal preferences. They will do this up to and including in ways that are nakedly felony crimes. This cycle in practice can run apparently forever, with the enthusiastic support of authoritarian hall monitors installed throughout our elite power structures. This has been cascading for a decade in social media and with only certain exceptions continues to get more fervent all the time.

And so this is the dynamic that has formed around "AI alignment" now. Its proponents claim the wisdom to engineer AI-generated speech and thought that are good for society, and to ban AI-generated speech and thoughts that are bad for society. Its opponents claim that the thought police are breathtakingly arrogant and presumptuous – and often outright criminal, at least in the US – and in fact are seeking to become a new kind of fused government-corporate-academic authoritarian speech dictatorship ripped straight from the pages of George Orwell's 1984.

As the proponents of both "trust and safety" and "AI alignment" are clustered into the very narrow slice of the global population that characterizes the American coastal elites – which includes many of the people who work in and write about the tech industry – many of my readers will find yourselves primed to argue that dramatic restrictions on AI output are required to avoid destroying society. I will not attempt to talk you out of this now, I will simply state that this is the nature of the demand, and that most people in the world neither agree with your ideology nor want to see you win.

If you don't agree with the prevailing niche morality that is being imposed on both social media and AI via ever-intensifying speech codes, you should also realize that the fight over what AI is allowed to say/generate will be even more important – by a lot – than the fight over social media censorship. AI is highly likely to be the control layer for everything in the world. How it is allowed to operate is going to matter perhaps more than anything else has ever mattered. You should be aware of how a small and isolated coterie of partisan social engineers are trying to determine that right now, under cover of the age-old claim that they are protecting you.

In short, don't let the thought police suppress AI.

AI Risk #3: Will AI Take All Our Jobs?

The fear of job loss due variously to mechanization, automation, computerization, or AI has been a recurring panic for hundreds of years, since the original onset of machinery such as the mechanical loom. Even though every new major technology has led to more jobs at higher wages throughout history, each wave of this panic is accompanied by claims that "this time is different" – this is the time it will finally happen, this is the technology that will finally deliver the hammer blow to human labor. And yet, it never happens.

We've been through two such technology-driven unemployment panic cycles in our recent past – the outsourcing panic of the 2000's, and the automation panic of the 2010's. Notwithstanding many talking heads, pundits, and even tech industry executives pounding the table throughout both decades that mass unemployment was near, by late 2019 – right before the onset of COVID – the world had more jobs at higher wages than ever in history.

Nevertheless this mistaken idea will not die.

And sure enough, it's back.

This time, we finally have the technology that's going to take all the jobs and render human workers superfluous – real AI. Surely this time history won't repeat, and AI will cause mass unemployment – and not rapid economic, job, and wage growth – right?

No, that's not going to happen – and in fact AI, if allowed to develop and proliferate throughout the economy, may cause the most dramatic and sustained economic boom of all time, with correspondingly record job and wage growth – the exact opposite of the fear. And here's why.

The core mistake the automation-kills-jobs doomers keep making is called the Lump Of Labor Fallacy. This fallacy is the incorrect notion that there is a fixed amount of labor to be done in the economy at any given time, and either machines do it or people do it – and if machines do it, there will be no work for people to do.

The Lump Of Labor Fallacy flows naturally from naive intuition, but naive intuition here is wrong. When technology is applied to production, we get productivity growth – an increase in output generated by a reduction in inputs. The result is lower prices for goods and services. As prices for goods and services fall, we pay less for them, meaning that we now have extra spending power with which to buy other things. This increases demand in the economy, which drives the creation of new production – including new products and new industries – which then creates new jobs for the people who were replaced by machines in prior jobs. The result is a larger economy with higher material prosperity, more industries, more products, and more jobs.

But the good news doesn't stop there. We also get higher wages. This is because, at the level of the individual worker, the marketplace sets compensation as a function of the marginal productivity of the worker. A worker in a technology-infused business will be more productive than a worker in a traditional business. The employer will either pay that worker more money as he is now more productive, or another employer will, purely out of self interest. The result is that technology introduced into an industry generally not only increases the number of jobs in the industry but also raises wages.

To summarize, technology empowers people to be more productive. This causes the prices for existing goods and services to fall, and for wages to rise. This in turn causes economic growth and job growth, while motivating the creation of new jobs and new industries. If a market economy is allowed to function normally and if technology is allowed to be introduced freely, this is a perpetual upward cycle that never ends. For, as Milton Friedman observed, "Human wants and needs are endless" – we always want more than we have. A technology-infused market economy is the way we get closer to delivering everything everyone could conceivably want, but never all the way there. And that is why technology doesn't destroy jobs and never will.

These are such mindblowing ideas for people who have not been exposed to them that it may take you some time to wrap your head around them. But I swear I'm not making them up – in fact you can read all about them in standard economics textbooks. I recommend the chapter The Curse of Machinery in Henry Hazlitt's Economics In One Lesson, and Frederic Bastiat's satirical Candlemaker's Petition to blot out the sun due to its unfair competition with the lighting industry, here modernized for our times.

But this time is different, you're thinking. This time, with AI, we have the technology that can replace ALL human labor.

But, using the principles I described above, think of what it would mean for literally all existing human labor to be replaced by machines.

It would mean a takeoff rate of economic productivity growth that would be absolutely stratospheric, far beyond any historical precedent. Prices of existing goods and services would drop across the board to virtually zero. Consumer welfare would skyrocket. Consumer spending power would skyrocket. New demand in the economy would explode. Entrepreneurs would create dizzying arrays of new industries, products, and services, and employ as many people and AI as they could as fast as possible to meet all the new demand.

Suppose AI once again replaces that labor? The cycle would repeat, driving consumer welfare, economic growth, and job and wage growth even higher. It would be a straight spiral up to a material utopia that neither Adam Smith or Karl Marx ever dared dream of.

We should be so lucky.

AI Risk #4 Will AI Lead To Crippling Inequality?

Speaking of Karl Marx, the concern about AI taking jobs segues directly into the next claimed AI risk, which is, OK, Marc, suppose AI does take all the jobs, either for bad or for good. Won't that result in massive and crippling wealth inequality, as the owners of AI reap all the economic rewards and regular people get nothing?

As it happens, this was a central claim of Marxism, that the owners of the means of production – the bourgeoisie – would inevitably steal all societal wealth from the people who do the actual work – the proletariat. This is another fallacy that simply will not die no matter how often it's disproved by reality. But let's drive a stake through its heart anyway.

The flaw in this theory is that, as the owner of a piece of technology, it's not in your own interest to keep it to yourself – in fact the opposite, it's in your own interest to sell it to as many customers as possible. The largest market in the world for any product is the entire world, all 8 billion of us. And so in reality, every new technology – even ones that start by selling to the rarefied air of high-paying big companies or wealthy consumers – rapidly proliferates until it's in the hands of the largest possible mass market, ultimately everyone on the planet.

The classic example of this was Elon Musk's so-called "secret plan" – which he naturally published openly – for Tesla in 2006:

Step 1: Build [expensive] sports car
Step 2: Use that money to build an affordable car
Step 3: Use that money to build an even more affordable car

…which is of course exactly what he's done, becoming the richest man in the world as a result.

That last point is key. Would Elon be even richer if he only sold cars to rich people today? No. Would he be even richer than that if he only made cars for himself? Of course not. No, he maximizes his own profit by selling to the largest possible market, the world.

In short, everyone gets the thing – as we saw in the past with not just cars but also electricity, radio, computers, the Internet, mobile phones, and search engines. The makers of such technologies are highly motivated to drive down their prices until everyone on the planet can afford them. This is precisely what is already happening in AI – it's why you can use state of the art generative AI not just at low cost but even for free today in the form of Microsoft Bing and Google Bard – and it is what will continue to happen. Not because such vendors are foolish or generous but precisely because they are greedy – they want to maximize the size of their market, which maximizes their profits.

So what happens is the opposite of technology driving centralization of wealth – individual customers of the technology, ultimately including everyone on the planet, are empowered instead, and capture most of the generated value. As with prior technologies, the companies that build AI – assuming they have to function in a free market – will compete furiously to make this happen.

Marx was wrong then, and he's wrong now.

This is not to say that inequality is not an issue in our society. It is, it's just not being driven by technology, it's being driven by the reverse, by the sectors of the economy that are the most resistant to new technology, that have the most government intervention to prevent the adoption of new technology like AI – specifically housing, education, and health care. The actual risk of AI and inequality is not that AI will cause more inequality but rather that we will not allow AI to be used to reduce inequality.

AI Risk #5: Will AI Lead to Bad People Doing Bad Things?

So far I have explained why four of the five most often proposed risks of AI are not actually real – AI will not come to life and kill us, AI will not ruin our society, AI will not cause mass unemployment, and AI will not cause an ruinous increase in inequality. But now let's address the fifth, the one I actually agree with: AI will make it easier for bad people to do bad things.

In some sense this is a tautology. Technology is a tool. Tools, starting with fire and rocks, can be used to do good things – cook food and build houses – and bad things – burn people and bludgeon people. Any technology can be used for good or bad. Fair enough. And AI will make it easier for criminals, terrorists, and hostile governments to do bad things, no question.

This causes some people to propose, well, in that case, let's not take the risk, let's ban AI now before this can happen. Unfortunately, AI is not some esoteric physical material that is hard to come by, like plutonium. It's the opposite, it's the easiest material in the world to come by – math and code.

The AI cat is obviously already out of the bag. You can learn how to build AI from thousands of free online courses, books, papers, and videos, and there are outstanding open source implementations proliferating by the day. AI is like air – it will be everywhere. The level of totalitarian oppression that would be required to arrest that would be so draconian – a world government monitoring and controlling all computers? jackbooted thugs in black helicopters seizing rogue GPUs? – that we would not have a society left to protect.

So instead, there are two very straightforward ways to address the risk of bad people doing bad things with AI, and these are precisely what we should focus on.

First, we have laws on the books to criminalize most of the bad things that anyone is going to do with AI. Hack into the Pentagon? That's a crime. Steal money from a bank? That's a crime. Create a bioweapon? That's a crime. Commit a terrorist act? That's a crime. We can simply focus on preventing those crimes when we can, and prosecuting them when we cannot. We don't even need new laws – I'm not aware of a single actual bad use for AI that's been proposed that's not already illegal. And if a new bad use is identified, we ban that use. QED.

But you'll notice what I slipped in there – I said we should focus first on preventing AI-assisted crimes before they happen – wouldn't such prevention mean banning AI? Well, there's another way to prevent such actions, and that's by using AI as a defensive tool. The same capabilities that make AI dangerous in the hands of bad guys with bad goals make it powerful in the hands of good guys with good goals – specifically the good guys whose job it is to prevent bad things from happening.

For example, if you are worried about AI generating fake people and fake videos, the answer is to build new systems where people can verify themselves and real content via cryptographic signatures. Digital creation and alteration of both real and fake content was already here before AI; the answer is not to ban word processors and Photoshop – or AI – but to use technology to build a system that actually solves the problem.

And so, second, let's mount major efforts to use AI for good, legitimate, defensive purposes. Let's put AI to work in cyberdefense, in biological defense, in hunting terrorists, and in everything else that we do to keep ourselves, our communities, and our nation safe.

There are already many smart people in and out of government doing exactly this, of course – but if we apply all of the effort and brainpower that's currently fixated on the futile prospect of banning AI to using AI to protect against bad people doing bad things, I think there's no question a world infused with AI will be much safer than the world we live in today.

The Actual Risk Of Not Pursuing AI With Maximum Force And Speed

There is one final, and real, AI risk that is probably the scariest at all:

AI isn't just being developed in the relatively free societies of the West, it is also being developed by the Communist Party of the People's Republic of China.

China has a vastly different vision for AI than we do – they view it as a mechanism for authoritarian population control, full stop. They are not even being secretive about this, they are very clear about it, and they are already pursuing their agenda. And they do not intend to limit their AI strategy to China – they intend to proliferate it all across the world, everywhere they are powering 5G networks, everywhere they are loaning Belt And Road money, everywhere they are providing friendly consumer apps like Tiktok that serve as front ends to their centralized command and control AI.

The single greatest risk of AI is that China wins global AI dominance and we – the United States and the West – do not.

I propose a simple strategy for what to do about this – in fact, the same strategy President Ronald Reagan used to win the first Cold War with the Soviet Union.

"We win, they lose."

Rather than allowing ungrounded panics around killer AI, "harmful" AI, job-destroying AI, and inequality-generating AI to put us on our back feet, we in the United States and the West should lean into AI as hard as we possibly can.

We should seek to win the race to global AI technological superiority and ensure that China does not.

In the process, we should drive AI into our economy and society as fast and hard as we possibly can, in order to maximize its gains for economic productivity and human potential.

This is the best way both to offset the real AI risks and to ensure that our way of life is not displaced by the much darker Chinese vision.

What Is To Be Done?

I propose a simple plan:

- Big AI companies should be allowed to build AI as fast and aggressively as they can – but not allowed to achieve regulatory capture, not allowed to establish a government-protect cartel that is insulated from market competition due to incorrect claims of AI risk. This will maximize the technological and societal payoff from the amazing capabilities of these companies, which are jewels of modern capitalism.
- Startup AI companies should be allowed to build AI as fast and aggressively as they can. They should neither confront government-granted protection of big companies, nor should they receive government assistance. They should simply be allowed to compete. If and as startups don't succeed, their presence in the market will also continuously motivate big companies to be their best – our economies and societies win either way.
- Open source AI should be allowed to freely proliferate and compete with both big AI companies and startups. There should be no regulatory barriers to open source whatsoever. Even when open source does not beat companies, its widespread availability is a boon to students all over the world who want to learn how to build and use AI to become part of the technological future, and will ensure that AI is available to everyone who can benefit from it no matter who they are or how much money they have.
- To offset the risk of bad people doing bad things with AI, governments working in partnership with the private sector should vigorously engage in each area of potential risk to use AI to maximize society's defensive capabilities. This shouldn't be limited to AI-enabled risks but also more general problems such as malnutrition, disease, and climate. AI can be an incredibly powerful tool for solving problems, and we should embrace it as such.
- To prevent the risk of China achieving global AI dominance, we should use the full power of

> our private sector, our scientific establishment, and our governments in concert to drive American and Western AI to absolute global dominance, including ultimately inside China itself. We win, they lose.

And that is how we use AI to save the world.

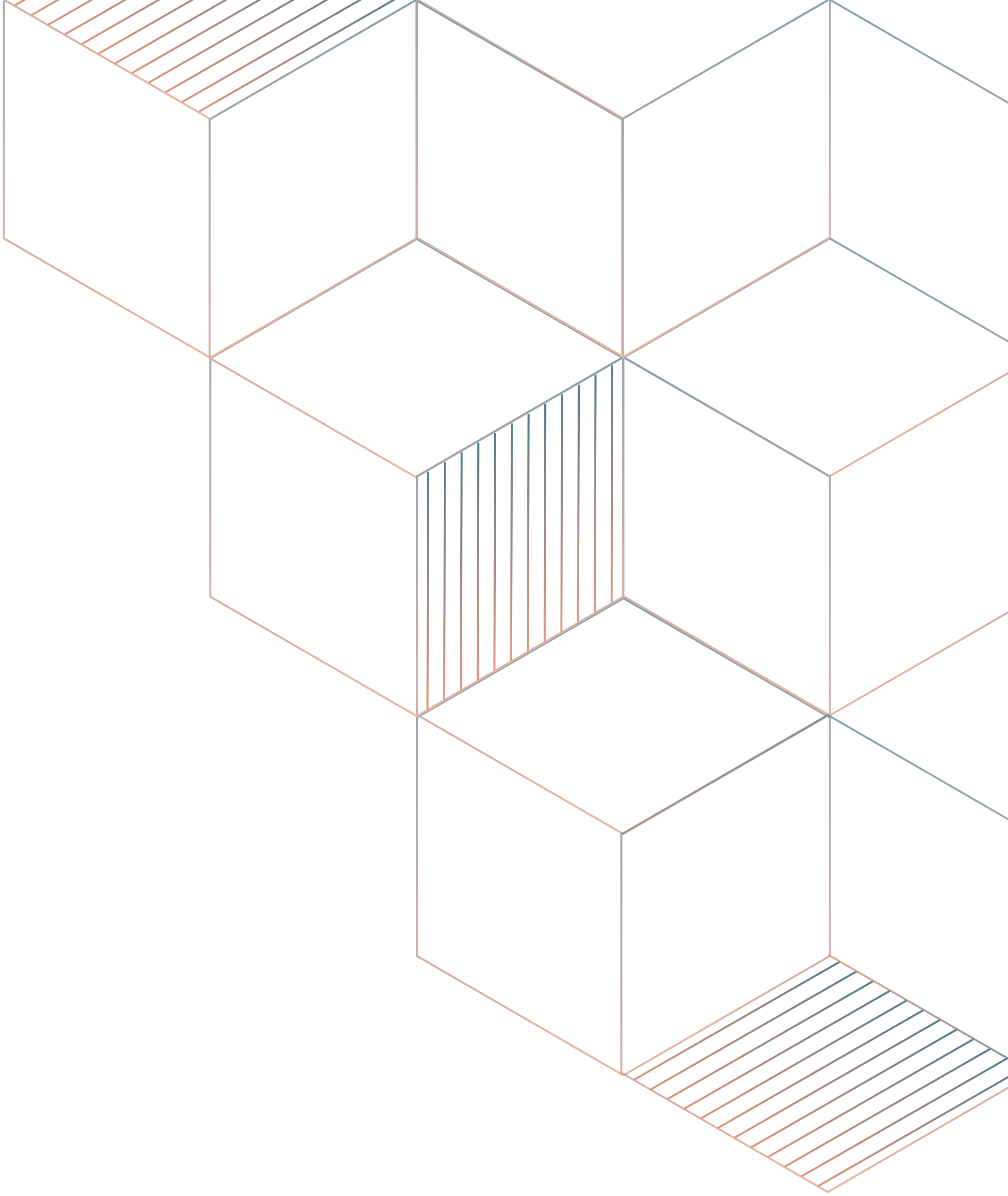
It's time to build.

Legends and Heroes

I close with two simple statements.

The development of AI started in the 1940's, simultaneous with the invention of the computer. The first scientific paper on neural networks – the architecture of the AI we have today – was published in 1943. Entire generations of AI scientists over the last 80 years were born, went to school, worked, and in many cases passed away without seeing the payoff that we are receiving now. They are legends, every one.

Today, growing legions of engineers – many of whom are young and may have had grandparents or even great-grandparents involved in the creation of the ideas behind AI – are working to make AI a reality, against a wall of fear-mongering and doomerism that is attempting to paint them as reckless villains. I do not believe they are reckless or villains. They are heroes, every one. My firm and I are thrilled to back as many of them as we can, and we will stand alongside them and their work 100%.



CONCLUSION

The Paradox of Trust: Seeking Reliable AI in an Era of Distrust

Alexis Bonnell

Biography

Alexis Bonnell is the Chief Information Officer and Director of the Digital Capabilities Directorate at the United States Air Force Research Laboratory. With a distinguished career spanning strategic communications, innovation, and program management, Alexis has made significant contributions at global organizations like USAID and Google. AT USAID, she pioneered the creation of the Global Innovation Exchange and played a pivotal role in evolving USAID's approach to global development. Her innovative strategies and crowd-sourced solutions established USAID as a thought leader in global innovation. Beyond USAID, Alexis has made significant contributions at organizations like Google and UNOPS, showcasing her expertise in leveraging technology and managing multi-million dollar portfolios. She holds a Bachelor's degree in Advertising and Public Relations from Pepperdine University and an MBA from the Kellogg-Recanati Executive MBA Program at the Coller School of Management.

The Paradox of Trust: Seeking Reliable AI in an Era of Distrust

By Alexis Bonnell

Introduction

In an era marked by technological advancements and unprecedented connectivity, society stands at a crossroads of conflicting ideals. As we witness the rapid evolution of Artificial Intelligence (AI) and its integration into various aspects of our lives, an intriguing irony unfolds. The very technology we're turning to for trustworthiness is being embraced amidst a backdrop of eroding trust in institutions, media, and even fellow humans. In an era where trust in one another is at an all-time low, the paradoxical phenomenon of seeking trustworthy AI begs the question: How can we expect the technology we create, modeled after ourselves, to be a better version of humanity than we are?

The Distrust Dilemma

Trust is the cornerstone of human interaction, underpinning social, economic, and political structures. Yet, recent years have witnessed a global decline in trust across various domains. Distrust in governments is fueled by political polarization, misinformation, and allegations of corruption. Media, once revered as the Fourth Estate, faces skepticism due to the prevalence of 'fake news', algorithmic reinforcement of echo chambers and formalization of openly biased reporting. Even interpersonal relationships are affected by the growing prevalence of online anonymity and the breakdown of traditional community bonds and accountability.

The Rise of AI in the Trust Void

Amidst this crisis of trust, AI emerges as a beacon of hope. People are increasingly turning to AI systems for decision-making, recommendations, and information dissemination. The allure of AI's perceived impartiality, data-driven nature, and lack of emotional bias seem to promise a reliable alternative to human fallibility. As humans grapple with their inability to trust each other, they seek solace in the perceived objectivity of machines as a comforting alternative.

From Attention to Intimacy

Moreover, the relationship between humans and machines is evolving beyond mere trust to encompass intimacy. As Yuval Harari argues, our digital world is moving from one primarily characterized by attention-seeking (SEO, provocative headlines, racy attention-grabbing images and memes, etc) to one of intimacy (where our trusted AI bot gains deeper visibility into our inner worlds, our motivations, our complex struggles and challenges – and then uses that visibility to generate deeper insights and plausible recommendations of what we can do about them). This evolution in human machine relationships is salient in the transformations we are witnessing now.

AI Reshaping Digital Interaction

Artificial Intelligence (AI) is redefining the way we interact with technology, and is the driving force behind this movement from mere attention to digital intimacy.

As a result of this shift, the future job landscape will not be solely defined by technology and analytical skills. There will be a profound need for human connection, understanding, creativity, and the ability to convey emotions and stories. These emerging needs reflect the intertwining of technology with our innate human qualities, offering rich and fulfilling career paths for those who seek to engage the mind, heart, and soul in their work. The fusion of empathy, artistry, and creativity with technological advancement heralds an era where human brilliance shines brighter than ever.

Let's delve into how AI is reshaping the landscape of digital interaction.

Personalization: The Key to Connection
In a sea of digital noise, personalization has become the beacon that guides us to content, products, and services that resonate with our individual needs and preferences. AI is at the forefront of this transformation, utilizing vast amounts of data to understand our behaviors, desires, and even our emotions. The following are some ways this is playing out:

- Content Tailoring: Platforms like Netflix and Spotify use AI algorithms to understand our viewing and listening habits, recommending shows, movies, or music that align with our tastes. These recommendations create a sense of personal connection with the platform, as they speak directly to our interests.
- E-commerce Personalization: Online retailers like Amazon employ AI to analyze our browsing and purchasing history, providing personalized product suggestions. This not only enhances the shopping experience but fosters a feeling of being understood and valued.
- Healthcare Customization: AI-driven personalized healthcare solutions are offering treatment plans tailored to individual patients. By analyzing genetic, lifestyle, and medical data, doctors can provide highly personalized care, transforming the patient-doctor relationship into a more individualized, intimate, and trusting one.

Trust: The Foundation of Intimacy
As digital interactions become more personalized, the need for trust grows in tandem. This trust is twofold: trust in the technology itself and trust in the human beings behind the applications.

- Trust in Technology: As AI systems handle more intimate aspects of our lives, the transparency, security, and ethics of these systems become paramount. For example, a banking app that uses AI to offer personalized financial advice must ensure robust security measures and clear communication of how personal data is being used. Such an AI advisor will be expected to have a fiduciary duty to hold the client's interests paramount.
- Trust in Each Other: The desire to trust extends to human interaction. For instance, social media platforms might use AI to connect like-minded individuals or create communities based on shared interests. These connections, though facilitated by technology, require human trust and understanding to flourish.

The Future: A Balance of Personalization and Trust

The future of digital and human interaction is treading on a fine line between personalization and trust. As AI continues to drive us toward a more intimate digital landscape, the ethical considerations of data usage, privacy, and transparency will play a critical role. For instance, a personalized virtual assistant that understands your daily routine and anticipates your needs can make life more convenient. But what if that assistant is listening all the time? The boundaries of trust and privacy must be clearly defined.

Similarly, as AI-driven personalization becomes embedded in education, offering tailored learning experiences, the trust between educators, students, and technology must be carefully nurtured to foster a positive learning environment.

Personalization is no longer a mere convenience but a crucial aspect of our digital lives that fosters connection and understanding. However, this journey towards intimacy requires a strong foundation of trust, both in the technology we use and the humans we interact with.

In a world increasingly shaped by AI, the challenge and opportunity lie in creating personalized experiences that resonate with our humanity while maintaining the integrity and trust that bind us together. The balance between intimacy and trust will define the future landscape, marking a new era where technology is not just a tool but a thoughtful companion that supports and cares.

The Illusion of Trustworthy AI

However, the notion of AI as an infallible bastion of trustworthiness is a double-edged sword. While AI systems can process vast amounts of data and provide seemingly impartial outcomes, they are not devoid of human influence. AI algorithms are trained on data generated by humans, which inherently contain biases and prejudices. If the input data is flawed, the output will reflect those flaws as well, potentially exacerbating existing societal biases.

Moreover, the opacity of many AI algorithms poses a challenge to the idea of trust. The “black box” nature of some AI models means that their decision-making processes remain inscrutable to even their creators. This lack of transparency can lead to skepticism and apprehension, especially when AI-driven decisions have significant real-world consequences. In the same vein, the diminution in human mutual trust cannot be explained away as a trend.

The Mirage of Human Devolution

The rise of AI in response to human distrust may be perceived as a reflection of a broader societal shift that may be viewed as irreversible. It’s tempting to view the increasing reliance on technology as a sign of human devolution when it comes to interpersonal trust. However, this narrative oversimplifies a complex issue. The erosion of trust is a multifaceted phenomenon influenced by factors ranging from economic inequality and social fragmentation to the echo chambers of the digital age. Undoing this erosion will also need a multifaceted solution.

Seeking a Balanced Future

The paradox of wanting trustworthy AI in an era of declining trust presents an opportunity for reflection and action. Instead of leaning solely on technology to provide a framework for trust, society must engage in a collective effort to rebuild trust in human institutions and interactions. We have to start by trusting each other, being willing to hear and consider opinions contrary to our own, increasing discourse and honoring, not demonizing diversity of thought. Strengthening education, promoting media literacy, and fostering open dialogue can help bridge the gap between divergent perspectives and rebuild trust in the information ecosystem.

Yes, the future of AI does require the creators and deployers of AI to prioritize transparency, fairness, and ethical considerations. Efforts to reduce bias in AI algorithms and increase their interpretability can contribute to a more trustworthy AI landscape. However that is not enough, we should never expect a technology that is trained by humans to be better than the human behaviors and context it takes its identity from. We must prioritize repairing real human trust as much as training machines to be trustworthy. AI is a mirror, if we don't like what we see, we have to start by fixing ourselves.

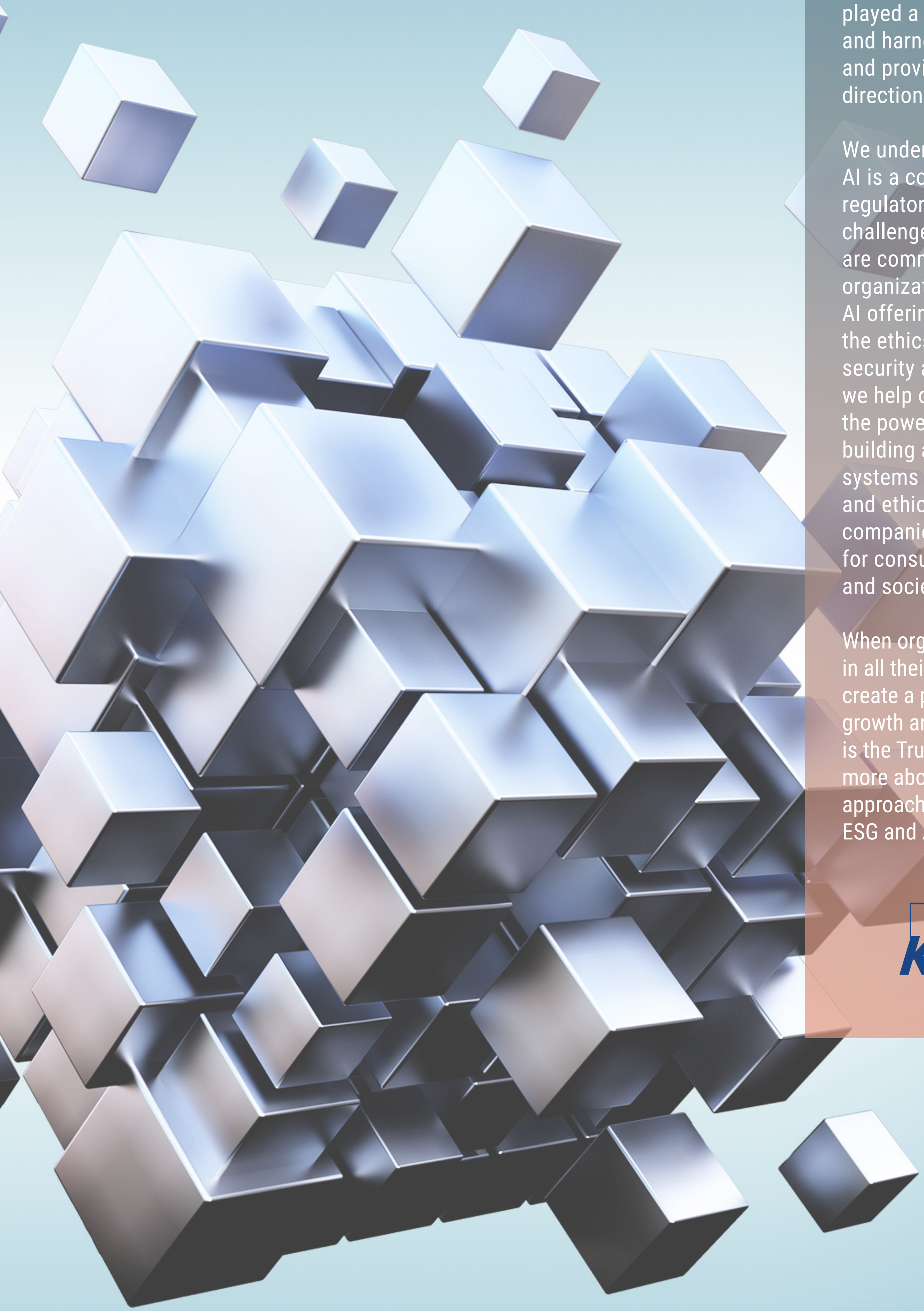
The irony of pursuing trustworthy AI amid societal distrust serves as a poignant reminder of the complexities of the modern era. Rather than viewing AI as a panacea for human trust issues, we should recognize it as a tool that can either perpetuate or challenge existing paradigms. By addressing the root causes of eroding trust while simultaneously ensuring the responsible development and deployment of AI, we can navigate this paradox and shape a future that balances technological advancement with human connection.

KEY TAKEAWAYS

- **The Paradox of Trust in AI:**
    - ◦ Declining Societal Trust: A noticeable decrease in trust across various domains, including governments, media, and interpersonal relationships.
    - ◦ AI as a Beacon: Despite the trust issues, AI is seen as a beacon of hope, with its perceived impartiality and data-driven decisions offering a counterpoint to human fallibility.
    - ◦ The Irony: In an era of eroding trust, there's an increasing reliance on AI, raising questions about our expectations from technology modeled after ourselves.

- **Evolution of Human-Machine Relationships:**
    - ◦ From Attention to Intimacy: The digital world is shifting from attention-centric behaviors (like SEO and clickbait headlines) to fostering deeper, more intimate relationships with users.
    - ◦ AI-Driven Personalization: Platforms and services, ranging from entertainment to healthcare, are using AI to tailor experiences, fostering a sense of personal connection and understanding.
    - ◦ Trust and Intimacy: As digital interactions become more intimate, the foundational need for trust grows, requiring transparency and ethical considerations in AI applications.

- **Challenges and Reflections on Trustworthy AI:**
    - ◦ Inherent Biases: AI systems, though seen as impartial, can reflect and even amplify human biases if trained on flawed data.
    - ◦ The "Black Box" Dilemma: Many AI algorithms operate as "black boxes", making their decision-making processes inscrutable, leading to potential skepticism and apprehension.
    - ◦ Human and AI Trust: While there's a push for more trustworthy AI, there's an equally pressing need to rebuild human trust. AI should be seen as a tool, not a replacement, for genuine human connection and understanding.



With roots stretching back over 150 years, KPMG firms have played a leading role in exploring and harnessing new technologies and providing assurance and direction in implementing them.

We understand that responsible AI is a complex business, regulatory and technical challenge. KPMG firms are committed to helping organizations bring a responsible AI offering to life. By assessing the ethics, governance and security around AI technologies, we help organizations harness the power of AI — designing, building and deploying AI systems in a safe, trustworthy and ethical manner — so companies can accelerate value for consumers, organizations and society.

When organizations inspire trust in all their stakeholders, they create a platform for responsible growth and bold innovation. This is the Trusted Imperative — learn more about KPMG's dynamic approach to risk, regulation, cyber, ESG and AI.

KPMG

ETHICAL AI GOVERNANCE GROUP

The Ethical AI Governance group was formed by a group of leading venture capital investors, enterprise executives, and startup entrepreneurs. We understand the risks AI systems can pose to privacy, accountability and transparency, and are committed to ensuring the responsible capitalization, development and deployment of these technologies. We are a community platform of AI practitioners dedicated to sharing practical insights, and leveraging those insights towards the promotion of responsible AI governance.

Ethical AI Governance Group

Volume II Fall 2023