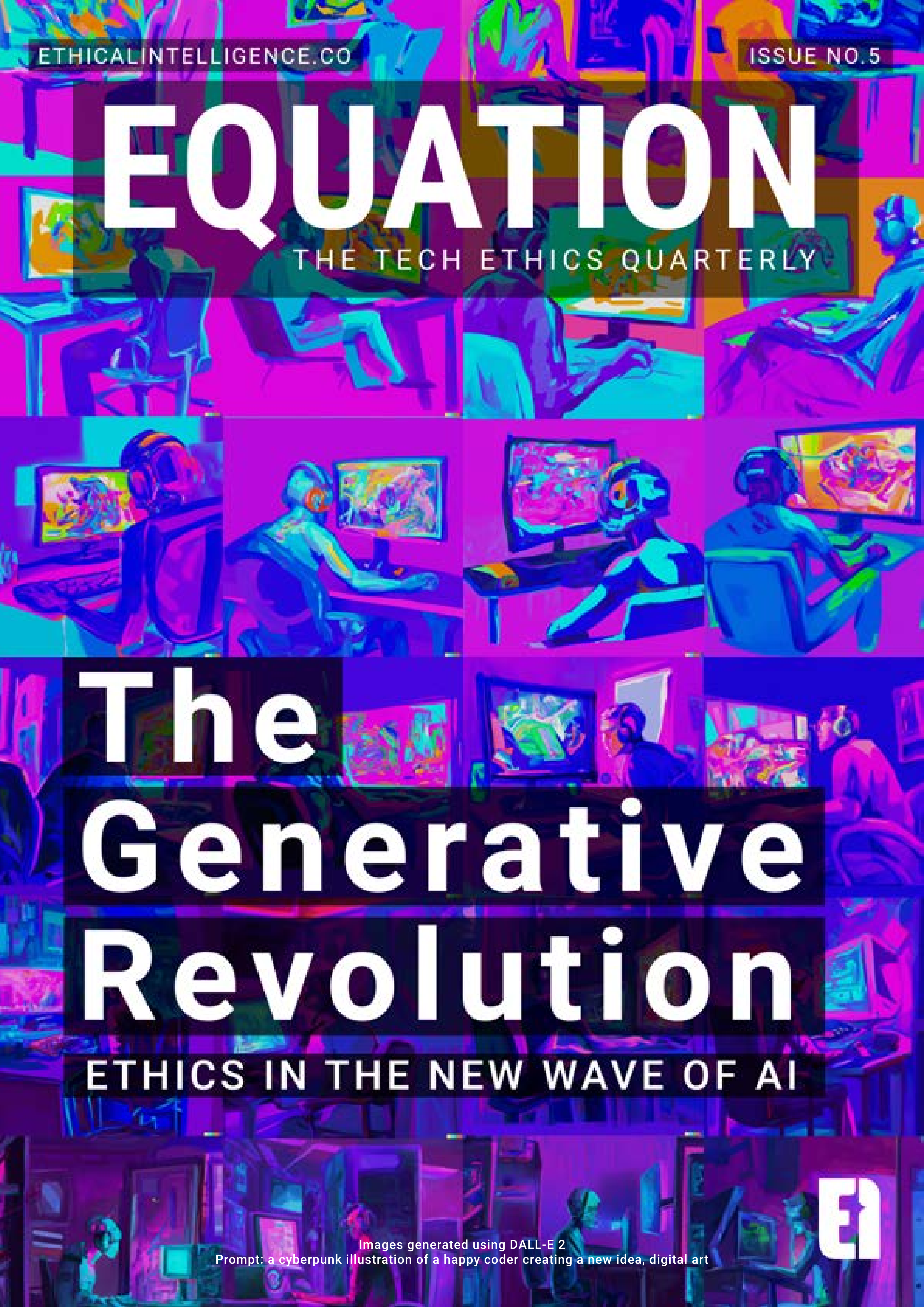
ETHICALINTELLIGENCE.CO

ISSUE NO.5

EQUATION

THE TECH ETHICS QUARTERLY



The Generative Revolution

ETHICS IN THE NEW WAVE OF AI

Images generated using DALL-E 2
Prompt: a cyberpunk illustration of a happy coder creating a new idea, digital art

Table of Contents

Subscribe for future issues delivered straight to your inbox

Image generated using DALL-E 2, prompt: generative ai artist, digital art

Dear Reader,

What do you fear most about generative AI?

Is it the privacy concerns about data sourcing and security? The bias embedded in the training processes? Or maybe the agency challenges in the questions around IP and patenting? All of these, and many more, are all very real and very valid risks - this is emerging technology, and there will always be risks when you introduce something new to the market.

But as those of you who know and work with me, I am not one to dwell on fear. As an ethicist, I will acknowledge and safeguard against the multitude of risks -but my focus has always been drawn to the opportunity we have as people to collaborate with the machines we are creating.

Generative AI applications like ChatGPT and Midjourney offer new and exciting avenues for growth. As you will find in this issue, we've embraced this new technology on a limited scale, using it to generate cover images and article titles - although the articles are exclusively human creations.

However, no matter the number of opportunities generative AI may open, with each new scandalous misuse of the technology, our mistrust continues to exponentially grow. We are quickly creating a deep-seated fear of generative AI, a fear that very well may cripple the technology altogether.

We will only ever see the full potential of generative AI actualized if we are able to trust how the technology is being built and used. And we will only ever be able to trust the technology if we ensure ethics has been embedded from the very beginning and that applications are being deployed responsibly.

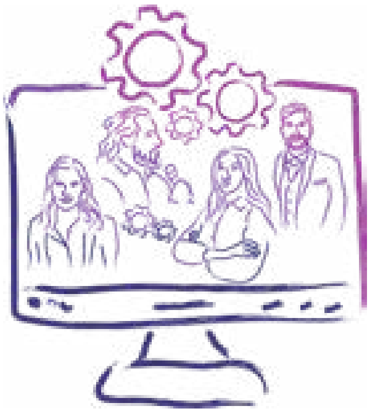
So, dear reader, if I had to answer my own question, my biggest fear of generative AI is that we will never see the full potential of the good we can do with it - simply because we have become too afraid of our own creation.

It is this fear that has inspired this issue of the EQUATION. Our hope is that, if the right information - such as that found in these pages - gets into the right hands, then we might be able to make even the smallest of dents in the course of this expanding technology.

Happy reading,

Special thanks to EAIGG for their partnership on the development of this issue.

EAIGG



We believe that ethics is the key to successful generative AI innovation and adoption.

Ethical Intelligence **partners** with industry leaders to empower your leadership, product design and tech teams to bring ethics to the forefront of your organization.

Together, let's remove the uncertainty that's been holding you back, and ensure **alignment** between your business objectives, values, and technology for a better tomorrow.

Interested in exploring what **responsible generative AI** could do for your organization?

*Let's talk - **info@ethicalintelligence.co***

Breaking Down the Basics

Understanding what is and is not Generative AI

Written by Jelmer van der Linde & Divyansh Agarwal

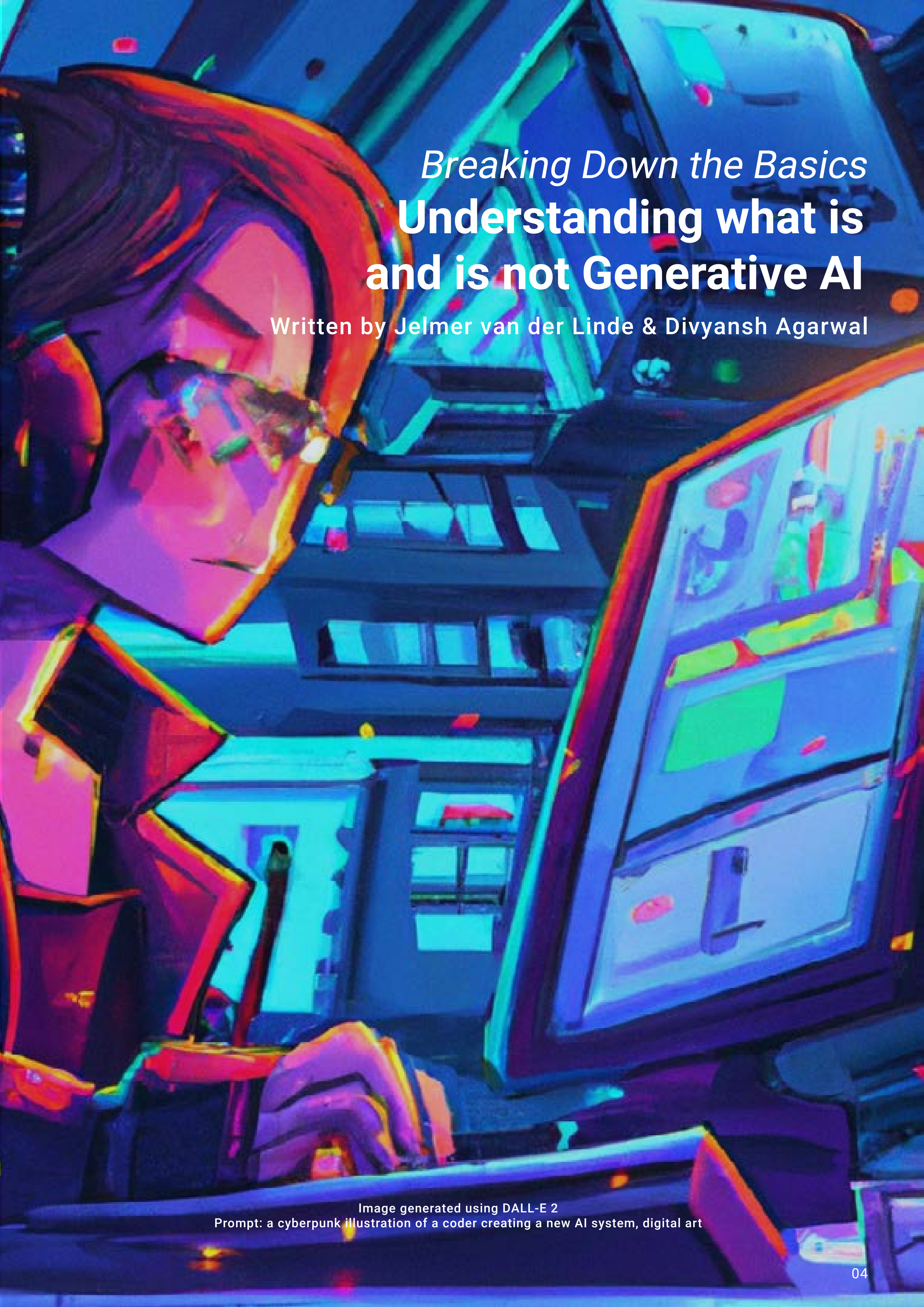


Image generated using DALL-E 2
Prompt: a cyberpunk illustration of a coder creating a new AI system, digital art

Generative AI is already the buzzword of 2023 in tech. The release of chatGPT by OpenAI has pushed AI 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 AI 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 AI scientists in the first place. At the same time, understanding how their performance is evaluated on different tasks and benchmarks, can help us to be more analytical in identifying the right use cases for generative AI. To develop this field in the right direction, we need to be able to distinguish between the systems that are built on top of Generative AI 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 AI is essentially a field of AI 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 AI model learn to understand simple natural language prompts and the expected response it should produce. Generative AI 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 AI 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 AI 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 AI 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.

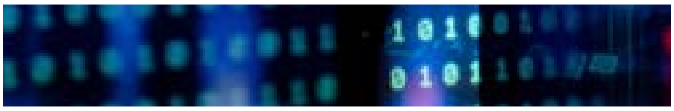
The different flavours of AI models that generate text, images, videos and multi-modal content, are equally impactful in their own respect. However, in order to get a better understanding of how these models are developed, let's dive deeper into text generation models like ChatGPT, which are more generally 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 AI 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?



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 - *a parameter is a unit used to classify AI systems, so the more parameters, the more complex the system.*

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 pre-trained LLMs will likely be able to for example perform ROT13, a simple substitution cipher, without an 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 realise 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 life-cycle 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.

Image generated using DALL-E 2
Prompt: programmer making generative AI in a compter lab, digital cyber punk

At A Glance

key takeaways from this article

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

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

Image generated using DALL-E 2
Prompt: impressionist painting of a group of lawyers

Exploring the Impact of Generative AI

on Intellectual Property Law and the Future of Artistic Expression

Written by Thibaut D’Hulst & Geoffrey Schaefer

Intellectual Property law is a bedrock of our legal system, yet it is surprisingly squishy terra firma, being both concrete in its protections and wildly interpretative. This has become a major challenge in the age of Generative AI. IP law covers many mediums but it's the painted work that presents the clearest example of this challenge, and one that we will return to throughout this article.

Consider a painter who is particularly inspired by the work of Jean Michel Basquiat and chooses to paint in a similar style. Over time, the artist becomes fairly competent and produces works that, while original, are increasingly representative of Basquiat's own portfolio. It becomes difficult to distinguish between the two artists' output. Is the inspired painter in violation of copyright law?

What if the artist is a Generative AI system? And the "inspiration" the AI took from Basquiat was it "seeing" the entirety of Basquiat's portfolio as part of its training data? Like the case above, the outputs are "original" but clearly inspired by the paintings of Basquiat. Did the AI system violate copyright law? Does Basquiat's estate have a case? We argue that in both cases the answer is "No." In the former, the artist may be jeered and his work written off as "derivative," but no serious lawyer would argue that a copyright violation had occurred. In the case of the Generative AI system, we argue that there are no factual or substantive differences; while the Generative AI system is a novel technology, its act of "creating" art is not.

This may seem like case closed. But this simulacrum of artistry presents much more challenging questions for society. What will it mean to live in a society whose predominant form of expression is synthetic, produced by the flipping of bits and not the human experience? And what role, if any, will IP law play with the growing use of Generative AI?

How Copyright Law Has Protected Artists Through Changes in Technology and Society

The history of copyright is closely connected with advances in technology and social norms. The invention of the printing press led to the first copyright law, Britain's 'Statute of Anne'. With this Statute, authors were given the exclusive right to authorize publications of their work. As technology advanced, new types of work qualified for copyright protection including sound recordings, audiovisual works (i.e., films), and eventually software. Of course, none of these existed when the Statute of Anne was enacted in 1710. Further technological inventions reduced the friction in the reproduction and disclosure of works, and in response to this "copyright," evolved into a bundle of exclusive rights to include: (i) the right to reproduce the work; (ii) the right to distribute the work; (iii) the right to publicly perform or display the work; and (iv) the right to make derivative works.

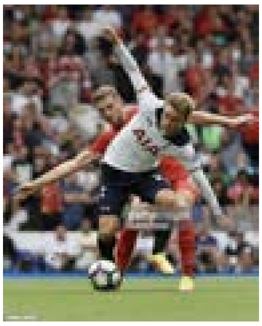
To qualify for copyright protection, works must be creative (or "original") and recorded in some form. But what do we mean by "originality"? In the E.U., the Court of Justice held that for a work to be original it must reflect the author's "own individual character"(1). Under U.S. law, a work is original if it is "independently created by its author and possesses some minimal degree of creativity" (2). Moreover, most legal systems require the work to be recorded in a fixed form to benefit from copyright protection. That means that an abstract idea or general concept (such as the character in a story or the style of an artist) cannot be protected by copyright.

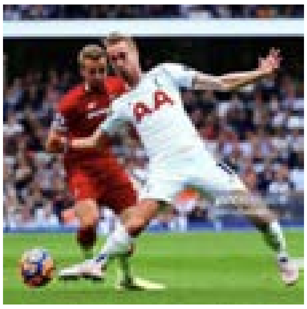
Generative AI programs that create text or images are very good at copying the style of an artist. This ability is not a new one; a skilled painter can produce a painting in the style of

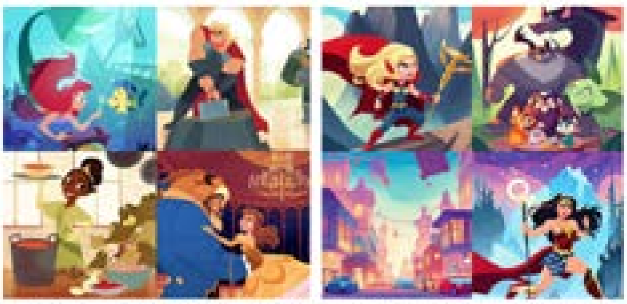
(1) European Court of Justice. (2009). Infopaq International A/S v. Danske Dagblades Forening (Case C-5/08). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:62008CJ0005&-from=EN

(2) US Supreme Court in the case of Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340 (1991)

Basquiat and a trained musician can compose a song redolent of the Beatles. In most cases, the new works would not infringe on the copyright of the original artist. However, the risk of infringement increases the more elements it adds from the original work. For instance, if our Basquiat-inspired artist adds additional stylistic hallmarks from one or more of his specific paintings - such as his signature composition or colour palettes - then the new work is more likely to infringe the copyright of the original work. A recent lawsuit brought by Getty Images against the AI company, Stable Diffusion, followed exactly this reasoning. On the left is the original image (©Getty images) and on the right is an AI generated image which Getty claims to infringe its copyright.







Artwork by Hollie Mengert (left) vs. images generated with Stable Diffusion DreamBooth in her style (right)

Even if AI-generated images do not strictly reproduce the original work, they could be seen as “derivative works”. (Derivative is both a pejorative term used to describe an artist’s work and a technical legal category. Under copyright law, “derivative works” include translations, arrangements, or a transformation from an original work such as the adaptation of a book into a movie. Derivative works generally require the consent of the original author, unless they can rely on a copyright exception.)

In U.S. case law, a major exception exists for the “fair use” of copyright works, which is an important concept in our discussion of Generative AI. In 1994 the U.S. Supreme Court had to decide whether the song “Pretty Woman” by the popular rap ensemble, 2 Live Crew, infringed the copyright of Roy Orbison’s classic, “Oh, Pretty Woman”. The Supreme Court held that there was no infringement “in light of the song’s parodic purpose and character, its transformative elements, and considerations of the potential for market substitution”. The extent to which a new work is “transformative” became an important element in determining whether the new work can rely on the fair use exception. The Supreme Court pointed out that “the goal of copyright, to promote science and the arts, is generally furthered by the creation of transformative works. Such works thus lie at the heart of the fair use doctrine’s guarantee of breathing space within the confines of copyright”. This can be understood as an affirmation that art must be allowed to be “inspired” by existing art, whether it’s produced by a human or an algorithm.

Case Closed? Not Quite…

Generative AI brings new challenges to copyright law. Courts will deal with these challenges by developing the existing case law and applying or adapting the criteria that we discussed above. Yet the precise nature of why Generative AI challenges copyright presents much broader societal concerns that we are only now beginning to understand.

Weighing the value of granting exclusive IP rights with the different incentive structures from granting their exceptions is not always an obvious calculus. While copyright was developed to protect the livelihood of creative professionals, the exceptions were introduced to accommodate other important rights and interests. The protection of fair use may cover freedom of expression (in the case of parody), freedom of information (for critique or reporting on current events). And the protection of “transformative” works balances the rights of different authors and their contributions in furthering their own artistic endeavors. Indeed, art very often develops from other art. An overly broad interpretation of derivative works could therefore stifle important stimulants of creativity.

Similarly, exclusive IP rights must be weighed against the societal interest in incentivizing new research and development. As a specific example, a recent E.U. Directive "on copyright and related rights in the Digital Single Market" introduced an exception to allow the mining of online content to facilitate innovative research and analysis. These decisions are often highly context-dependent and can fluctuate with the details of the case at hand.

What Should We Protect in a World of Ubiquitous Production?

The history of IP law is at an inflection point. The principal question concerns how IP law should evolve to find a new and acceptable equilibrium between human creators and AI generators. We may find it necessary, if not desirable, to rethink the entire nature of what should be protected and why, and how AI systems themselves - as they become closer to autonomous agents - should be considered in this legal paradigm.

We're now seeing the first tangible symptoms of Generative AI's disruption in the creative sector. Sci-fi magazine Clarkesworld no longer accepts submissions for publications after they were spammed by AI-generated stories. Copyright advocates warn that allowing copyright protections for AI-generated content could open the door for copyright trolls. For example, if a Generative AI published all possible arrangements in music, then any new hit single would immediately be hit by a copyright claim. While these trolling strategies are technically possible today, Generative AI makes them practically so by significantly reducing the friction of arranging, recording, and publishing music at scale. Each of these actions previously required copious time and expertise, but they can now be performed by the push of a button.

This disruption is likely to produce a new equilibrium that may have detrimental effects on creative industries. We have seen this before. Take the music industry as an exemplar. Beginning with peer-to-peer file sharing, music all of a sudden became widely available for free. Neither extant laws or the subsequent court cases could stop the ensuing infringement of copyright protections on a massive scale. But the industry eventually evolved its business model, providing musical content through either a monthly subscription or free with advertising. Generative AI will only turbocharge similar transitions in other industries. Creativity is resilient by nature, but creative industries often suffer before they flourish.

New Roles for a New World

What role should the artist play in a world simultaneously understood, created, and influenced by Generative AI? Is there a difference between the creativity inherent in a human's artistic expression and that of a machine? Will we care about the process behind a work of art, or simply its aesthetic beauty? One thing to consider is that Generative AI is now, at the very least, a source of creative output. Artists can choose to compete with these systems, or leverage them as tools in their own creative process. We emphasize the competitive dynamic here intentionally, as companies will increasingly use Generative AI systems to produce marketing copy, images, graphic design, and other artistic elements in the course of business. Whether prompted by a human or not, Generative AI systems are now creative actors in their own right and companies will use these systems when it makes creative and economic sense. This will force artists to either level-up their own work, or co-opt these systems to produce art that's greater than the sum of their parts.

What role should the development teams behind these Generative AI systems play? Soon, the capabilities of these systems – largely due to their generalizability – will become untethered from the original intent they were imputed with. Development teams will have little control over what these systems are able to produce and how. As such, any "IP controls" deemed necessary and/or legally mandated will need to be built into a system during its initial design and training. Even still, this may not prevent, say, the production of a photographic image like in the Getty case described above.

Adobe offers another approach. They recently released a suite of tools for graphic designers and other artists to introduce Generative AI capabilities into their workflow, synthesizing the artist and the machine. This is a clear example of a development team using the power of these systems to aid artists instead of producing tools that could undermine their economic viability. Not all Generative AI systems can or should be built to serve artists and other producers of original work; but nor should they be neglected as stakeholders in the design and use of these systems.

Conclusion

The biggest question of all is what this means for society. The immense scale that Generative AI systems operate means that we're nearing a scenario in which society's collective artistic output will be produced by machines and not humans. In this scenario, the world that we know and engage with will be, by definition, artificial. The stories that we read, the paintings that we view, the movies that we consume will all be fragments of training data. Art will be new but familiar, original but fake. And the more that we engage with it, the more training data we'll feed into the Generative AI systems and power their parrot-like production back to us. This may herald the end of artistry. Or it may spark a second Renaissance. Us humans are never content with stasis. We loathe sameness. A culture that's sourced principally by Generative AI promises to produce just this, however, arresting the engines of our creativity. But artists abhor banality. A synthetic world is anathema to them. Creatives will forever be inflamed to create.

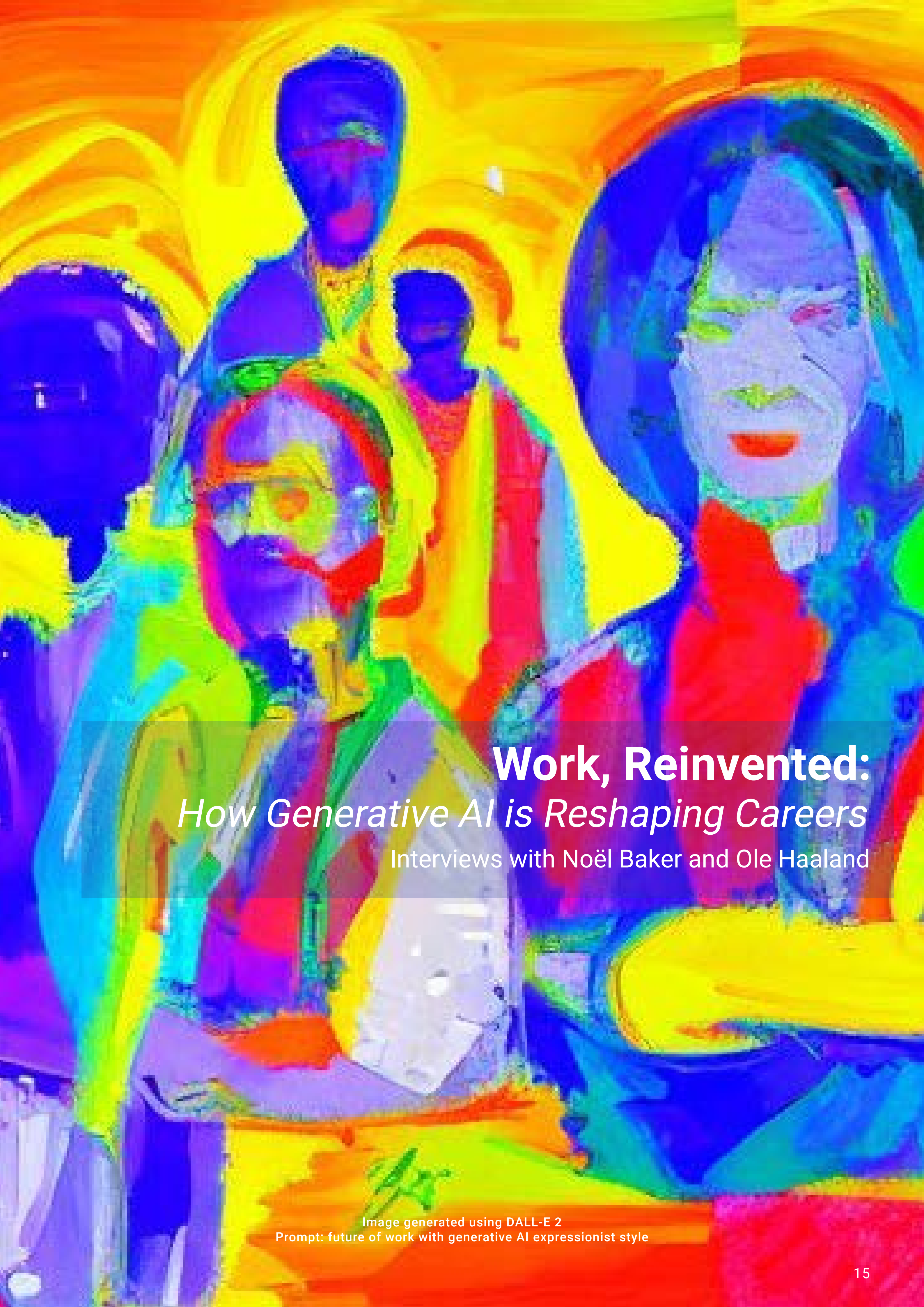
What role will IP law play in all of this? No meaningful role at all. Sure, there will be a slew of new court cases with novel but fundamentally similar arguments to one's that have been made in every generation since the Statute of Anne. But the history of copyright law shows that the march of artistic progress continues untrammeled. In some ways, then, this is not a matter of law at all; this is simply us creatives raging against the machines.

Image generated using DALL-E 2
Prompt: impressionist painting of a group of lawyers

At A Glance

key takeaways from this article

- While Generative AI is a novel technology, its process of "seeing" and "producing" art is functionally the same as a human artist.
- IP law has always struggled to provide stringent protections for artists without stifling art itself. Generative AI is but the latest challenge to this equilibrium.
- The bigger societal question is what our world will look like when the predominant form of artistic expression is produced by the flipping of bits and not the human experience.



Work, Reinvented:

How Generative AI is Reshaping Careers

Interviews with Noël Baker and Ole Haaland

Image generated using DALL-E 2
Prompt: future of work with generative AI expressionist style

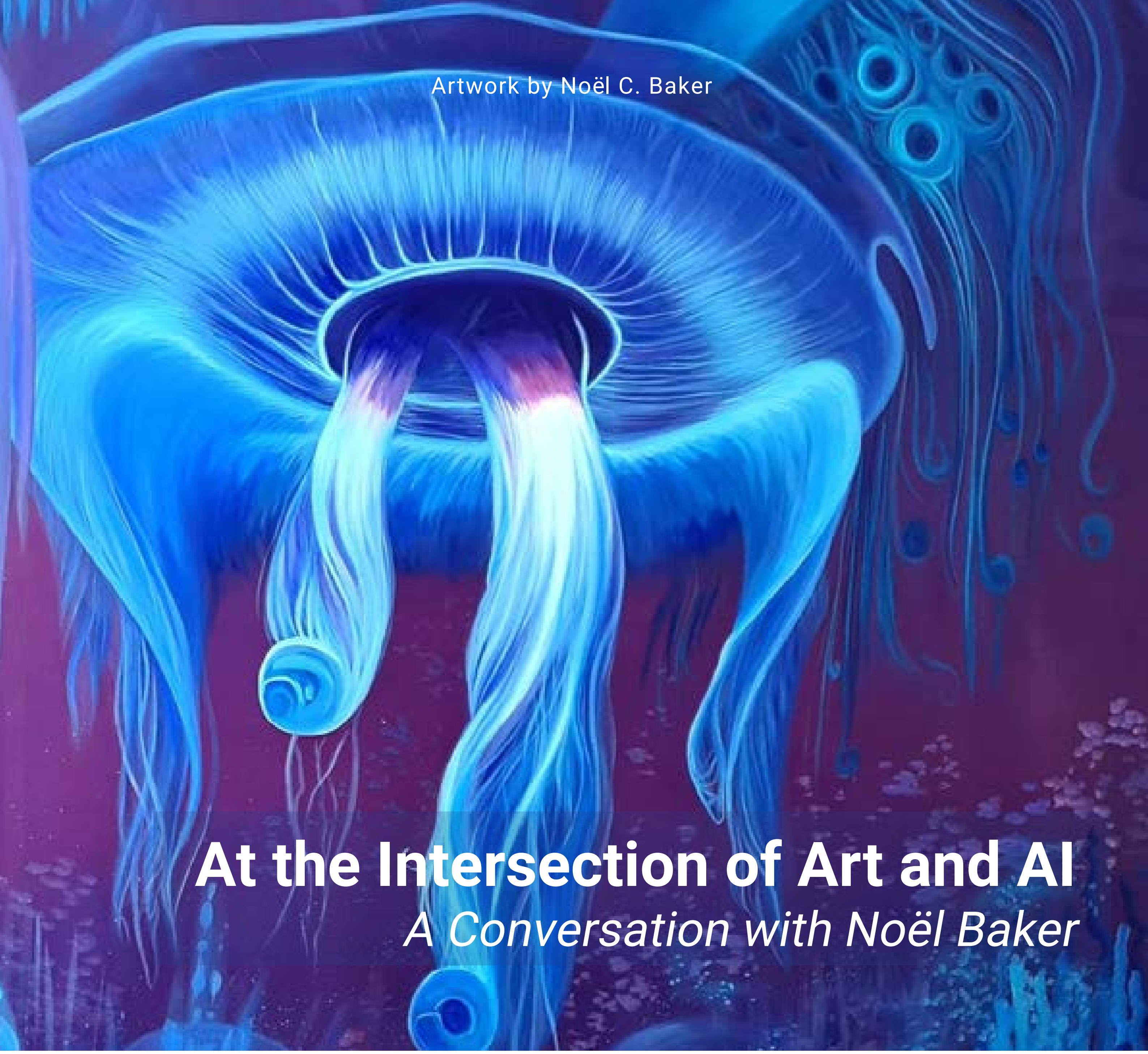
The ongoing development and proliferation of generative AI has led to both trepidation and excitement about its impacts on work and the workplace. As employers begin to evaluate the myriad of opportunities presented by the integration of generative AI systems into their working practices, the increasing expectation - in some areas already beginning to be realised - that AI will fundamentally alter working life and disrupt a multitude of tasks, roles, and careers across a range of sectors has understandably led to fears among employees. Perhaps the most widespread, and the most frightening, is the fear of replacement; as when any technological advance promises to automate tasks that fell previously under the exclusive domain of human activity, the new integration of these technologies will fundamentally alter the way we go about our day-to-day tasks. This incoming change has, understandably, fed into fears surrounding the future of work.

However, while many remain cautious, there are optimists who are inviting this change with welcome arms. The automation of basic routines and tasks has the potential to increase productivity, and leave more time for creative and critical thinking skills which are unlikely to fully replicated by AI systems. So, while certain skills and roles may become redundant, other - potentially more fulfilling - outlets will emerge, which will continue to require our human touch. As we adapt and integrate generative AI into our work we will need ask: what shifts will emerge as certain skills and roles are automated? What steps can we take to mitigate the negative consequences of this transition? And, most importantly, how will we use our time, skills, and human touch to remodel our purpose in the workplace?

The following interviews tackle fundamental questions surrounding the future of work, the integration of generative AI into the workplace and the potential for fostering creativity through generative AI. We interviewed the artist - Noël Baker, and engineer Ole Haaland, who are optimistic about the opportunities unfolding with generative AI. Instead of trepidation, Ole and Noël ask us to be open to these technologies and grow alongside them.

Noël C. Baker, PhD.
Artist, Climate Scientist

Ole Haaland
Robotics Engineer



Artwork by Noël C. Baker

At the Intersection of Art and AI

A Conversation with Noël Baker

How are you incorporating generative AI into your artwork? How did you first start thinking about AI in relation to art?

I see generative AI as a tool, rather than the artist, just like any tool that's come about in the art world or to humanity in general. It's not like I'm using what AI generates, and that's the art and that's done. What I do is I use it as a tool to help me imagine, to give me different possibilities and take inspiration from. I use it to make a painting of my own creation; at no point is it a direct copy of something else. So it's almost a collaboration more than it is a generation of art, because the end product is inspired by what I ask the AI to produce.

I've done a few art exhibitions, and I'm working on one big expo that's coming up from April to June in Brussels, called "Seas & Oceans" (pictured above). And one of my main involvements with generative AI is the huge art piece I'm doing for this expo.

As a climate scientist, I wanted to talk about my relationship with environmental grief, and the experiences that I've had as a scientist being exposed to the changing environments, and the impact that humanity has on the degradation of the natural world. Through art, I imagine a world after climate change has occurred, after humanity has done the worst that we can do, to the point that humanity has killed itself off and most of the species on the planet. And all that's left in the seas and oceans are species that could survive the impacts of climate change on the oceans which are warmer waters, higher dissolved carbon, basically a more acidic, warmer environment. I was trying to imagine what this would look like. A world after all of the species that we know and are familiar with have died off, and new species have evolved from the resilient ones that were able to survive climate change. So I wanted creatures that don't currently exist, that have never existed on the planet.

So, I asked AI to generate creatures for me. The generative AI creates its own interpretation, where things are changed and warped. They are a little bit off and different, but still realistic enough that you can imagine them being real creatures. It ended up being the perfect collaboration, and I used these bizarre AI output creatures to populate my underwater world.

I must have spent 30-40 hours playing with the AI generator to get all these different creatures. And I went through hundreds and hundreds of different AI generated images to find the ones that I liked. And then I combined, and mixed, and just created this whole imagined new world.

A lot of it was just wanting to play and have fun, which is one of the joys of art; I see it as an opportunity to try and expand our horizons. Isn't that really what art is about?

How do you see the role of generative AI in the art community?

AI used to be a tool that artists didn't pay much attention to. But now it's becoming so big that it has begun to threaten certain art communities. As an example, a friend of mine is a purely digital artist who doesn't do any traditional painting on canvas; everything is done with a computer on a graphic design program. She feels deeply threatened by AI because everything she does can be reproduced to some reasonable-quality amount by AI.

This is a very serious threat for digital artists because their art can be used as training sets for the AI, and an artist's style can be reproduced with exceptionally good fidelity to the point that you cannot distinguish what the AI produces from the artist's original work.

I can understand why she feels threatened by the proliferation of generative AI. But my perspective is very different, and I think this is because I'm not a purely digital artist. Most of my art is paint on canvas. So, I don't feel the same existential threat from AI. But I also feel like my approach and my perspective towards AI is more of the way a scientist might approach it. I'm not worried about AI because I use it purely as a tool.

Artwork by Noël C. Baker

Do you think there a way to misuse generative AI as a tool in the artworld?

Well, it's an interesting question from the overall perspective of art. I like to look at it in the way we think of art history. So, consider a Renaissance artwork. It's sometimes taken to be the 'purest' form of art. They look at a subject and create incredible lifelike portraits. But even they used a lot of tools and tricks. So even back during the golden age of the Renaissance, there is evidence that they used what technology they had at the time: optical instruments, mirrors, even projection technology to trace the images of what they were seeing directly onto the canvas.

Of course, they have wonderful skills to make the portrait realistic, but they're skipping a lot of the painful process of getting the proportions.

So, what is pure art that's untouched by technology?

At every step of the way along the path of art all the way up to now, with any tool that has been introduced, some people are going to say, oh, this is going to ruin art, this is going to ruin the whole industry, it's going to ruin what makes an artist pure. There are always going to be people who are afraid of it or who say that this is the end and we're never going to have purity like we once did. And while I understand there are serious problems with AI and art, I think it's largely confined to the digital artist's perspective.

In my opinion, generative AI is like any other technology: it can be used in the right way and in the wrong way, and the user should be given the knowledge and ability to make that decision to use it the best way they can.

Artwork by Noël C. Baker

What kind of opportunities do you see in using generative AI as a tool in your art? What are you excited about around having this onset of generative imagery for artists?

Accessibility is definitely the first one. Going back to my friends, the Renaissance artists. It used to be that if you wanted to become an artist, you would either have to be rich with a lot of free time on your hands, or you'd have to study for years. It was only something that was available to those few students who could study under those few masters.

Now that accessibility is open to anybody with an internet connection, which is just tremendous. I mean, the fact that anyone can pick up a pencil and paper is already a wonderful thing about art that it is accessible to anybody. But the ability to train and learn new things is becoming more and more open. I recently heard a story about someone who generated a children's book entirely with AI images and generated text. He got a big backlash from the art and literature community because he made a whole children's book which was purely generated, and he did it in a weekend.

But at the same time, he had a vision, he found the tools, and he produced something that's of pretty good quality. I mean, that's kind of cool from a scientist's perspective. Like, how fun is it that you can now create something wonderful with a little bit of time and energy, and a lot of inspiration and passion. I find that wonderful and democratic.

Generative AI creates art in seconds, whereas a human artist will take hours or even months to create a single piece. Is there any inherent value, beyond the final artwork, that can be attributed to the time it takes for an artist to create the artwork?

On one hand, this guy produced a book within a weekend, which is incredibly fast. But artists were very quick to poke holes in the art that was generated. Hands, for example, are something that generative AI absolutely fails at. If you've ever tried to generate a handshake, it's like there are 12 different fingers on each hand, or there are like three hands that are all grabbing different arms and elbows. And it's just a monstrous mess.

It's interesting because hands are also something that are traditionally very challenging even for pure traditional artists.

Even with producing art with generative AI, there's a learning curve. To get an output that is actually decent requires quite a bit of time. It's not a small task, and it's not as straightforward as you might think. You can't just push a button and have what you want to come out. It still takes work and effort.

I almost see that as a beautiful parallel between AI being in its infancy and artists learning from a young age how to do art, because hands are something that both fail at so spectacularly and take a lot of time to get right.

With AI as it is now, it's a useful tool for some things, but it's not nearly skilled enough to reproduce the Mona Lisa or any classical art piece that took many years of mastery, technique and perfection to reach.

So, time certainly still has value.

A couple months ago, I found someone on Instagram who had these beautiful watercolour paintings. They were so dreamy and strange; the details were like something straight out of a dream. Things are placed in ways you never would think. The people were all styled in this very strange, not modern style. I was just fascinated. So, I followed this person, and I was baffled because they were producing tons of art - like one or two pieces per day - and just getting lots of followers. And I couldn't understand how they were making such amazing watercolor paintings so quickly. And then I looked a little bit further into their descriptions, and somewhere down there at the very bottom, it said AI generation; they had no other non-generated artwork. This was their art. It was all produced with AI.

So was it misleading? I don't know. I was misled, but maybe I just wasn't reading closely enough the first time.

Artwork by Noël C. Baker

When you realized those pieces were generated images, was your perspective or appreciation of the artwork altered? Do you think it matters how transparent artists are about their use of tools in creating art?

My first reaction was, oh, well, this isn't nearly as good as I originally thought. I was very impressed by it considering what I thought in the beginning, which was that it's a pure watercolor artwork. I was very impressed at their imagination and use of colors and details. It was just mind-blowing when I found out it was generated by AI; my first emotional reaction was disappointment.

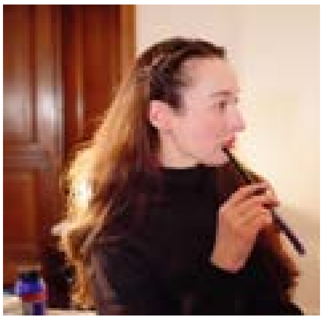
And then my second one was that I wanted to try to recreate that! I wanted to figure out exactly which AI tool they used and exactly which inputs they used. I wanted to get exactly what they made, and through experimentation, I actually got pretty close, but not exact. So that was satisfying to me; it was almost like a game to try to figure out how they did it.

But does that mean it has less value? Again, instinctively, my first reaction is to say yes, because the amount of time it takes and the amount of work they put in to generate AI is a lot less than someone who was doing it traditionally. But does it have less value to the people who see it, and are inspired by it, and enjoy looking at beautiful watercolor paintings, and maybe buy one and put it on their wall? I'm not sure that it would.

What would you advise another artist curious about generative AI to do?

Play, have fun, explore, be creative, but make sure that what you produce is coming from you, make sure that it's your creation in the end. That's what I would tell an artist.

Follow Noël's work on instagram: **@noel.c.baker**



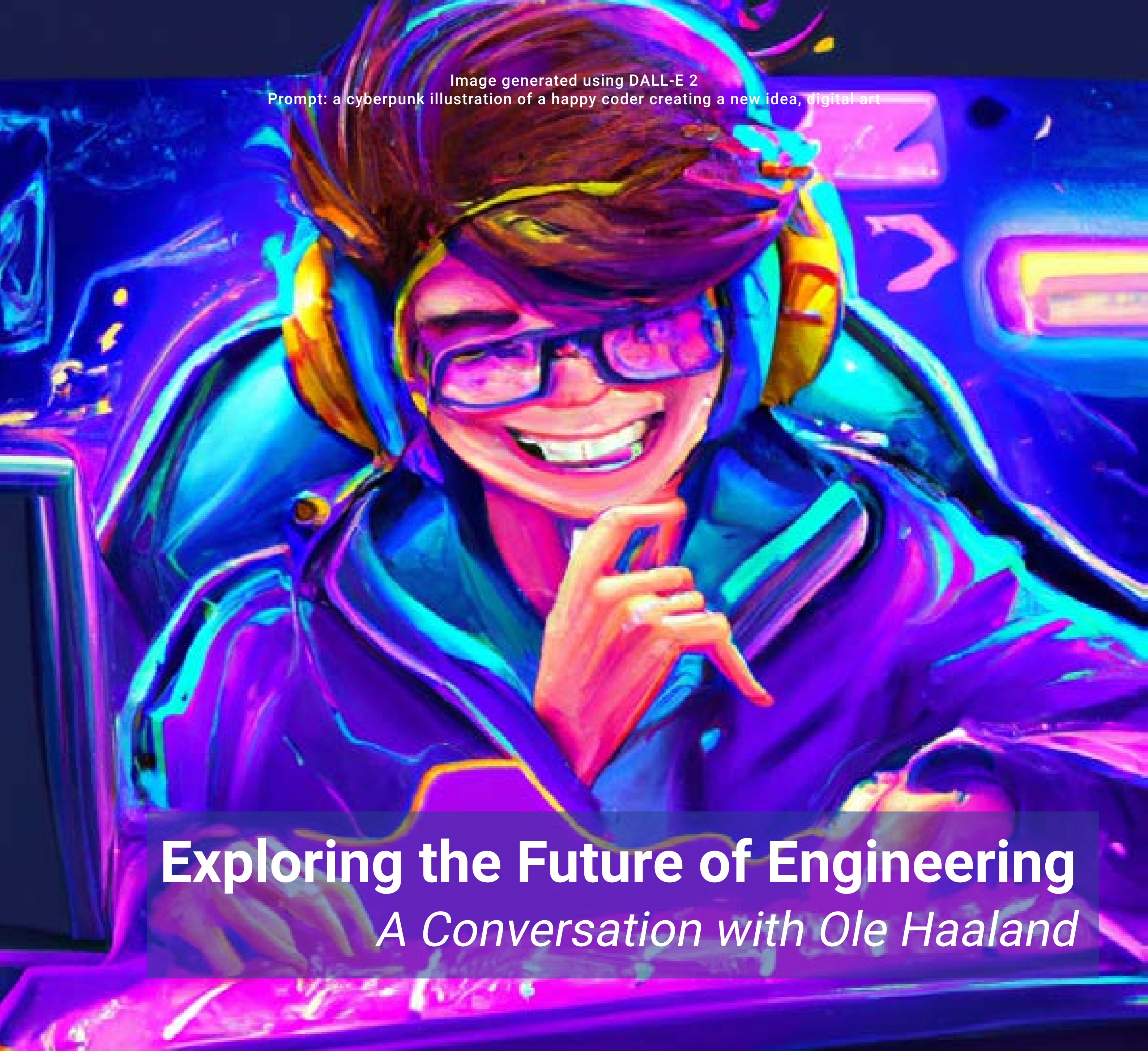


Image generated using DALL-E 2
Prompt: a cyberpunk illustration of a happy coder creating a new idea, digital art

Exploring the Future of Engineering

A Conversation with Ole Haaland

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

A couple of years ago I came across this generative AI tool called GitHub Co-Pilot - it's a generative AI 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, 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 AI 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 AI 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 use them.

How do you see the role of engineers changing in response to the increased use of generative AI?

I think there are two core aspects to think about here. First, many of the skills we care about aren't relevant to our work yet. Second, creativity and system thinking will become our most important skills.

So number one is that certain skills might become less needed in the future. At school, spelling and grammar are taught as essential skills in life. Doing well at school, and even passing our exams required us to be good at it. But this might change when generative AI is fully integrated into our lives. Knowing these small nitty gritty details will probably be much less important in the future.

And this goes for legal professionals and programming too. You don't need perfect spelling because you can just put broken text into Chat GPT and get perfectly corrected text. So many of these skills that you'd need a high level of precision for might not be needed anymore.

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 ChatGTP'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 ChatGTP 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, ChatGTP is incredibly averse to conflict and would therefore be a great moderator. So in this way, generative AI might actually aid collaboration.

What other positive implications do you see with generative AI 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 AI?

My biggest worry about generative AI 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 AI will not exactly slow this trend, but rather help get us even more hooked. What happens when AI-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 AI 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-specialised, 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 ChatGTP 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.

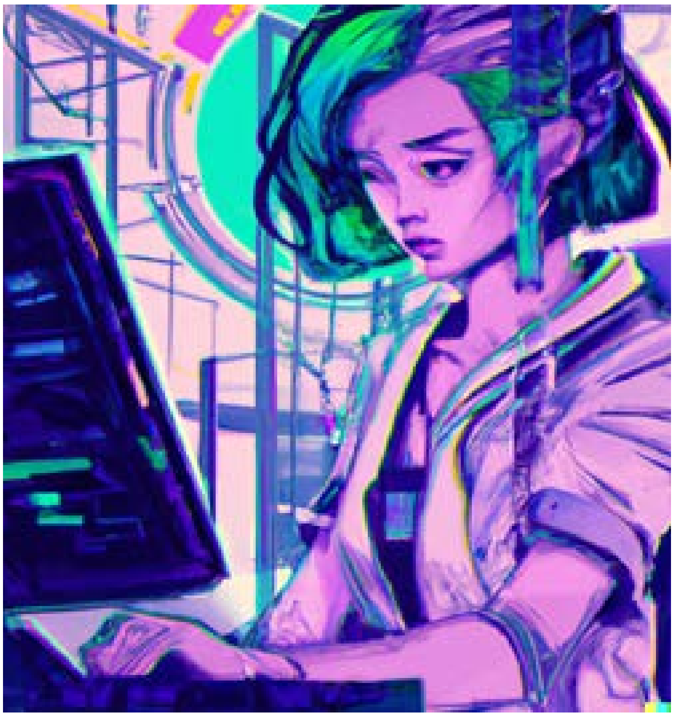


Image generated using DALL-E 2
Prompt: a cyberpunk illustration of a coder creating a new AI system, digital art

At A Glance

key takeaways from this article

- Our roles, skills, and workplaces are likely to change considerably as generative AI 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 AI 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 AI 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 sceptical about generative AI. 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 AI 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.

Discover what Ethics-as-a-Service can do for you.

As your **partner** in the development of ethical AI, we're here to work with you and your teams to find the ethics solution that's best for you.

Training

Empower your people with the culture, knowledge, guidance they need to foster a culture of responsible innovation and informed development.

Strategy

Gain the confidence and clarity you need to push boundaries by understanding the specific ethical challenges and opportunities relevant to your technological success.

Governance

Give your teams clear direction on how to build and use responsible technology by upgrading critical internal processes to reflect your values.

Interested in exploring what **responsible generative AI** could do for your organization?

*Let's talk - **info@ethicalintelligence.co***

At What Cost?

The Impact of Business Models on the Ethical Landscape of Generative AI

Written by Alexandra Crew & Matthew Douglas

Image generated using DALL-E 2
Prompt: business models in technology, graphic style with no text

Generative AI will transform every industry. Of this you are likely already aware. And yet, it seems that amongst all of the discussion about the capabilities and potential of new tools such as ChatGPT, Stable Diffusion, and others, there is significantly less discussion about the impact of the business models associated with them. As powerful as these technologies are, their availability and the impact they have is largely dependent on the business decisions made by their creators regarding how these tools are released into the world. As new, industry-specific Generative AI tools become more prevalent, how will their creators bring them to market?

The lack of significant discussion about business models is most likely tied to the fact that these early tools have leveraged free or freemium business models. This business model has been favored given the creators primary concern of driving rapid growth of a broad user base. But this era of open access and free tools will not last forever. As the shift to industry-specific toolsets and narrower user bases occurs we will see new ethical questions arising for the developers of such Generative AI tools. The impact of decisions such as pricing and revenue strategy, delivery channels, and even customer targeting will all come with their own set of ethical outcomes that will be just as important as understanding the technology itself.

Life or Death in AI

There is possibly no more high-stakes example of this than in the world of healthcare. Replicating the years of professional training, knowledge, and experience of physicians and other medical professionals to discern truth in incredibly complex medical scenarios is not a simple task. And yet, the potential for AI to supplement the expertise of medical professionals is an exciting one.

You may have seen recent news about one experiment that showed ChatGPT's ability to perform at or near the pass threshold on the United States Medical Licensing Exam (USMLE), or perhaps you've heard about the progress Google is making with its medical LLM, Med-PaLM 2, passing the USMLE with a score of 85%. While these LLM technologies present an exciting potential to transform patient care they also create a bevy of new ethical concerns. Medical consultation chatbots therefore present an intriguing case study through which to explore the power of business models in limiting or exacerbating such concerns.

Before moving forward, we want to be absolutely clear that neither of the previously mentioned AI solutions are ready for use in clinical settings, and we do not currently recommend the use of LLMs in medical settings. While their potential is exciting, there are also very serious concerns about hallucination and its impact on potentially life or death scenarios which must be sufficiently addressed from an ethical perspective before any such tool should be made available for use.

The At-Home AI Medical Consultation

Companies exploring the possibility of creating and launching a medical consultation LLM can dramatically shift both business and ethical outcomes by adjusting business model inputs such as revenue models, service delivery channels, and the target profile of their end user. To prove this point we will consider two potential business models for the same LLM powered medical consultation chatbot.

Imagine a company called SaludAIble has created a technology that can chat with patients, ask questions to uncover symptoms, gather relevant screening data, and make a prediction as to whether intervention by a medical professional is needed without the individual patient ever needing to leave their home. As a user, I don't have to think twice about whether I feel 'ill enough' to seek medical attention. I can simply open an application on my phone and get some quick advice on whether I am worrying too much, or I should actually schedule that appointment to go see my doctor. How could SaludAIble profitably yet ethically bring this offering to market?

Business Model 1
The Open-Access Dilemma

- Open-access via publicly accessible website
- Revenue generated from advertising through the website
- No direct involvement from Healthcare Providers

If SaludAIble were to provide the technology for free, directly to the general public through a website, they could generate large advertising revenues without having to ever sell directly to healthcare providers. offering a chatbot for free would certainly increase access to important medical advice on a broad scale and would be especially helpful in increasing access for those facing a high cost or a long commute to a doctor's office. It may also improve speed of triaging, decrease visits to healthcare facilities and in turn reduce strain on overburdened facilities.

It is also likely that increased access will lead to more conditions being caught earlier that may have otherwise been noticed later or may have gone undiagnosed.

However, this increase in access comes with its own concerns. For example, if the threshold at which the chatbot recommends a patient to seek medical advice is too low, this high volume of new patients may lead to unnecessary visits and a waste of healthcare resources. Also, as access is increased and the group of users becomes larger and more heterogeneous, the developer will need to ensure that the data utilized is representative of this broad user base, in order to mitigate bias and enable accuracy across sub-groups.

Without the involvement of a healthcare organization to provide important contextual data as inputs for each user interaction, errors and inaccuracies will also be more common. If a user is seeking medical advice but the chatbot has no access to the patient's health record, then the accuracy of the chatbot will be limited. The chatbot would not have access to crucial information regarding the patient's history.

Accountability is another ethical and medico-legal consideration that takes a different form without the direct presence of a healthcare organization as we will see in the second business model. With the dyad of the patient and maker of the chatbot, it must be presumed that any issues of liability will fall solely on the creator of the chatbot, should the chatbot provide any incorrect medical advice to a patient. This is not a trivial matter when dealing with life or death consequences.

It is also important to consider issues of accountability arising from displaying advertisements alongside medical advice. Users may become confused, or even steered towards the purchase of some product or service because of the timing of an advertisement displayed while chatting with the chatbot. Imagine you seek advice from this tool because you have a high fever, a cough, and severe muscle pain and upon inputting your symptoms you notice an advertisement for flu medication. Anyone would be more likely to purchase this flu medication in such a scenario, and there are both ethical and legal issues to consider should that medication not be appropriate. It is extremely important to consider how and when ads are displayed as well as what data is drawn from to determine which ads are shown to a user.

On top of all of this, it is also worth considering that the broader the user base for the platform, the more rapidly any ethical risk is likely to scale.

Business Model 2
Providers in the Loop

- Accessed via app or website of specific healthcare organization
- Revenue generated by licensing platform to healthcare organizations
- Healthcare organizations provide additional data inputs to model including patient medical history

How do the ethical dynamics change when access to such a tool is licensed directly to and provided to patients by a healthcare organization (e.g. health system, payer, provider group) instead of being freely available online?

The patient could still access this AI consultation with similar ease and simplicity, but instead of navigating to a general webpage, they may now need to login into an app provided by their healthcare provider or insurer. With this change in service delivery channel, the target user profile has also changed to focus on end users whose experience can be enriched by using medical records stored with their trusted provider. With the patient’s health record as important contextual information, the chatbot will be able to offer more tailored, accurate advice and thereby solve the generalization problem of the first business model.

There are further benefits to adding a provider to the loop when it comes to issues of accountability and liability. By securing buy-in and drawing on the expertise of healthcare organizations as customers of the platform, issues of accuracy would be more proactively addressed during proof of concept engagements with these customers. Risks of liability will also be reduced because the service will be provided to fewer total patients allowing accuracy to be improved iteratively with a smaller user group.

While the additional oversight of healthcare organizations will offer some benefits, adding an additional party that might be responsible for negative effects on patients complicates questions of accountability. This may open up potential accountability gaps which will be crucial to address.

Another point of difference from our first model is that access to the benefits of this tool will be restricted solely to patients who have some form of affiliation with a healthcare organization who can afford to pay for it. This will lead to a widening of healthcare disparities in which healthcare organizations with significant resources are able to provide this service for their patients while organizations and patient communities with fewer resources may not be able to do so. Think of rural communities with already limited access to medical services in comparison to wealthier metropolitan areas dense with healthcare offerings. These communities who already have more limited access to healthcare are often the communities who would stand to benefit the most from such a tool. This is a tradeoff that cannot be overlooked.

Values Informing Tradeoffs

Our hope in writing this article is that we inspire leaders, innovators, and you - our readers - to get involved in thinking about and discussing not only the ethical impacts of generative AI itself, but also to see and understand the power of business models in shaping an ethical landscape.

Aligning your business model to your company values is a crucial step to ensuring positive impact. In the end, any company exploring AI must start by mapping their company values to their choice of business model. As a first step, we encourage you to take the time to clearly map out your company values and the implications of each component of your current or planned business model on ethical factors including access, accuracy, accountability, transparency, and bias. Once you've clearly defined how your business model and company values support or hinder ethical outcomes, you'll be prepared to start making decisions to improve in the areas that are right for your organization, your customers, employees, and all other stakeholders.

Image generated using DALL-E 2
Prompt: rocks balanced on the sand, wes anderson film style

At A Glance

key takeaways from this article

- Changing a business model can be as impactful, if not more, on ethical outcomes than a technology itself.
- Designing an ethical business model is a dynamic process dependent on many variable factors.
- Aligning your business model to your company values is a crucial step to ensuring positive impact.

Building Sustainable Monetization Strategies for Open Source Generative AI

Written by Yaron Zakai-Or

Image generated using DALL-E 2
Prompt: generative ai artist, digital art

Open source companies have "officially" been around for 24 years, but the practice of releasing free software with its code has been around since the early 1990s. Open source has proven business models, growth tactics, licensing schemas, and huge adoption. One cannot imagine the world without Android, Linux, Git, MySQL, Kubernetes, React to name a few.

In the last couple of years, Generative AI has turned from a niche technology into one of the most impactful technologies around. It creates a new world of opportunities, developing at light speed, but still needs to mature in terms of technological robustness, ethics, and a comprehensive set of ML platforms.

Generative AI is in its initial hyper growth phase and many believe it will have a significant impact on both the workforce and economy. In a recent report from Goldman Sachs, the report's authors, which include Chief Economist Jan Hatzius, said "roughly two-thirds of jobs in the U.S. and Europe are exposed to some degree of AI automation while generative AI could replace up to 25% of current employment, or some 300 million full-time jobs, but will improve GDP by 7%".

The monetization opportunity for open source companies in generative AI is large, based on what has already been achieved with traditional software and AI; open source revenues have grown from $11.4 billion in 2017 to as high as $32.95 billion in 2022. According to the MarketsandMarkets Report, the growth forecast is an 18.2% CAGR, which means that it will reach $50 billion by 2026, compared to $21.7 billion in 2021.

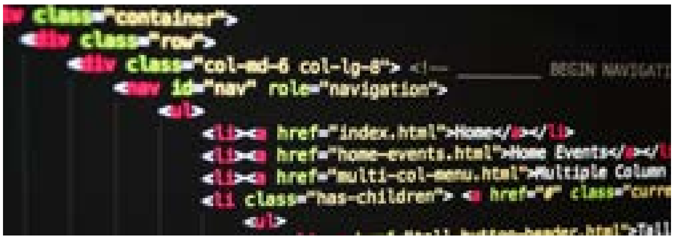
The significant initial resource investment to develop Generative AI models and applications brings a new monetization challenge to open source. By discussing the key monetization opportunities, we find that there is no one size fits all solution. Instead, businesses will need to evaluate the right strategy for them by performing fast experimentation on multiple strategies and perform further experimentations within each monetization strategy.

Generative AI and Open Source

Startups and established companies that are taking part in developing Generative AI solutions can be put into several key categories: ML PLatforms, chatbots, text, video and image, search, code, and others. There are already key open source projects and companies in most of these areas. One of the most exciting developments is GPT-J and GPT-Neo from Hugging Face, an open source alternative to GPT3. There are examples of open source companies in search (Qdrant, Deepset), ML Platform (Jina AI, Rubbrband), Synthetic Data (DataCebo), and most likely a lot more on the way.

As open source generative AI products continue to gain popularity, their creators are facing three standard challenges that come with building and maintaining an open source project. They are all important considerations which companies seeking to build successful open source generative AI products will need to bear in mind

- **Generating a thriving and sustainable community** of users and contributors around their open source project. This involves building a community of developers who are passionate about using the product and some who are willing to contribute their time and expertise to the project.
- **Developing a sustainable monetization model.** In the sections below, we will evaluate several options for monetizing Generative AI open source projects, building the business models for these projects is a complex and interesting challenge.
- **Protecting their intellectual property** (IP). One of the challenges in open source projects is around IP, releasing all your codebase to the public means that your competitors have an easy way to know exactly what you are doing. This is a good place to get an initial introduction to Open source and IP.



```
 class="container">
  <div class="row">
    <div class="col-md-6 col-lg-8"> <!--  ________  BEGIN NAVIGATI
      <nav id="nav" role="navigation">
        <ul>
          <li><a href="index.html">Home</a></li>
          <li><a href="home-events.html">Home Events</a></
          <li><a href="multi-col-menu.html">Multiple Colum
          <li class="has-children"> <a href="#" class="curre
            <ul>
```

The Uniqueness of Generative AI

Generative AI is bringing about a new open source paradigm because it involves more than just code. The training of a generative AI model relies on multiple interdependent components, including:

1. The model's code (consisting of its components and configuration)
2. Training data
3. Computing power used during the training process, which can be very demanding

For every open source software project, its creators should think about protecting their IP and keeping their technology defendable. This is also applicable to Generative AI. Protecting software IP is not a straightforward task, but it involves well known tactics. Every well established law firm with a tech practice will be able to advise which licensing fits your case, be it GPL, AGPL, MIT license, Apache license, or others.

Generative AI open source projects create a new challenge: one needs to defend the IP in the training data and the investment in computing. Your competitor can take the same data, apply the same set of hardware servers and reproduce the same results. On the licensing side it's a good idea to start from the common license types for datasets on kaggle. That means you have to put more effort into creating a sustainable and defendable licensing and monetization strategy. We will focus on the monetization strategy here, but it's highly recommended to look into the licensing side carefully before releasing your open source project.

There is a key ethical advantage to releasing your Generative AI models and applications as open source: the transparency of the data and the code enables an increase in the ability to assess important qualities of these models and applications, such as fairness and bias. Another potential ethical benefit is that the use of open source Generative AI models can foster diversity and inclusion by lowering the barriers to entry and participation for individuals and communities who may otherwise lack the resources or skills to access or create generative AI applications and models.

Open Source Monetization

In order to provide a more accurate framework, we will have to separate Generative AI solutions into two groups that differ in where their key IP exists:

1. **Training-heavy solutions,** such as GPT-J, which uses significant computational power. In this category we will also have the applications that fine tune an LLM with large data sets. Note that the data here is of equal importance to the model, which implies that the IP is beyond the model/code.
2. **Applications and infrastructure solutions** that fit into the "traditional" model: most of the investment and the IP is in code.

We will review four monetization models: Open Core, Professional Services and Support, SaaS and Marketplaces. For each of these models, we will relate to the above classification of Generative AI solutions: Training-heavy solutions vs. Applications and infrastructure solutions.

Open Core

Open core means that the company is releasing its core software functionality as an open source project and builds a software product with extra features around it. The idea here is to make the open-source software ("core software") become the standard among more advanced software developers and charge for the easier deployment and advanced functionality that fits organizations with the need to move fast at the expense of paying for the software product. This has become one of the leading monetization models for open source companies due to its simplicity in pricing and maintainability.

Moreover, an interesting trend of patterns emerge for open core companies, where their commercial offering is built around extending the open core with one or more of the below patterns:

Ease-of-use

- UX, Collaboration tools
- This is one of the best ways to create differentiation between open-sources and paid products. It involves creating a polished user experience (UX) and user interface (UI). In many cases, creating a collaboration layer on top of your open source product enables your customers to support business processes inside and outside the company. Integration with Slack is a good example for such a collaboration capability.

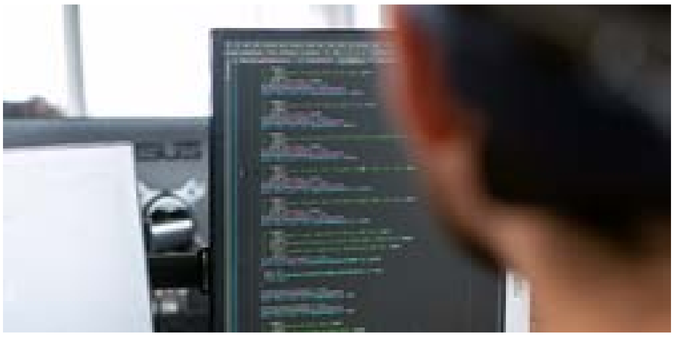
Enterprise

- Scalability, Security, RBAC (Role-based access control) and Integrations
- Enterprise implementations usually involve requirements on the IT side, including: security, scalability and interoperability with other enterprise applications.

Solutions

- Use-case specific functionality
- Enabling your core functionality to shine in specific verticals (e.g. Healthcare, Automotive, etc.) or to a specific set of users (e.g. CMOs in large corporates), you can create an appealing offering that will convince them to switch from open source to paid.

Open core is the model for recent open source success stories, including Confluent, Elastic, and Gitlab. Open core fits all flavors of Generative AI solutions. One of the advantages of this monetization model is that it keeps the ethical advantages of open source Generative AI models in its core/open source form, while it enables monetization.



Professional Services and Support

Early open-source models were often built on professional services. In this model, customers pay for support and implementation/consulting efforts. It is appealing to customers that are looking for an accountable business partner that will give them a source to rely on for bug fixes, implementation architecture and customization. Many of the early open source success stories were based on this model, including Red Hat. The Red Hat case is interesting as they today have multiple monetization paths, including Professional services and support as well as Open Core. One should experiment with multiple monetization options, similar to what RedHat did. Within each of the models, there is additional space for experimentation, such as: what SLAs for support do you get with each price tier?

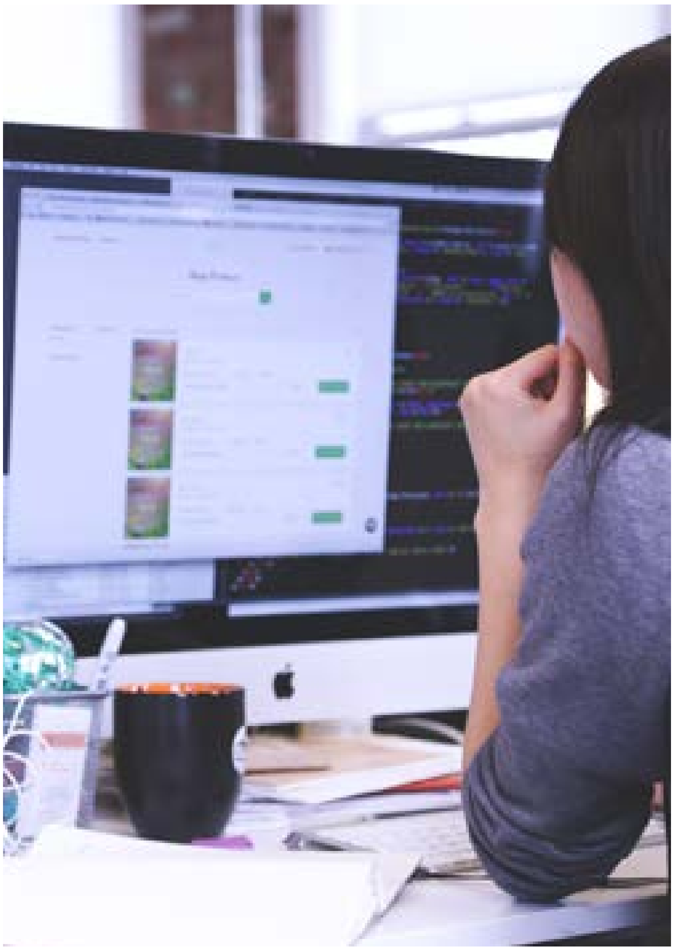
This monetization strategy was popular during the early days of open source, and it still exists. But one has to carefully evaluate if it fits their business, as it makes it hard to build a moat around your solution with this option when your IP is in training the large language model. However, it is hard to build a real moat around your business with professional services. As a result, in most cases this wouldn't be a great way to monetize your investment in open source for Generative AI.

Code-based solutions can still use this business model, assuming their knowledge of the solution space is wide enough and can create a moat. For training-heavy solutions, with the extra significant investment that you put into

Generative AI, my recommendation is to avoid this monetization strategy.

As this model does not truly support Generative AI models projects, it does not enable the true ethical advantages of such models.

In general, if you are doing this for a startup, stay away from this monetization strategy, as it doesn't fit well with how most VCs measure the potential success of startups. Professional services and support can be a small component of your revenues, but the best would be to sell your solution using a relatively predictable business model, such as SaaS.



SaaS

SaaS is a natural way to monetize open source projects. The idea is simple: host your open source project in the cloud, provide your customers with security, privacy, scalability, SLA and support and voila, you have a product. This has proven to work for Redis, MongoDB, Gitlab, Elastic, and confluent. There are a few benefits to this monetization strategy:

1. Combine with open core: you can provide additional features in your SaaS offering that are not available in the open source project. This is the case for Elastic and Confluent.
2. Modifying your open source license to prevent your competition from offering the exact same SaaS offering based on your open source project.
3. Offering "Enterprise" editions which offer higher level of service, dedicated infrastructure, and other enterprise-grade features, which makes this a hybrid with "Professional Service and Support"

One of the key advantages of this monetization model is around ethical AI: as it enables its creators to keep their IP protected, but provides transparency to the code, model and the data it is the best of all worlds, assuming that SaaS offering open source project is indeed a viable model and/or application independently.

Enterprise options can go all the way to supporting an on-premises installation, which sets a higher support cost, but would significantly improve your margins. Margins for the SaaS environment vary quite a bit and depend on the solution's data storage and compute needs, and of course on the competitiveness of the environment. But even with imperfect SaaS margins, the right mix of SaaS and Enterprise deals can create a healthy margin.

Experimentation here is key and it can be around which features are supported only in your SaaS product, which enterprise editions are offered and what features differentiate their tiers. There is no recipe for success here and one needs to try and adopt via measured experimentation.

Marketplaces

Creating a marketplace around your open-source solution is a compelling and off-the-beaten-track monetization strategy. Two well-known marketplaces are built on top of open source solutions including Android, likely the most successful open source story out there. Google doesn't break out revenues for their Play store, there is a good base number as a reference. As part of a lawsuit in 2021, it was disclosed that Google Play revenues were [$11.2 billion] in 2019. The second example is the

GitHub Marketplace, built around the open-source Git.

While this is a less well-proven path (in terms of the number of companies that are using it), it has the potential to be a real moat. This is especially relevant if a trustworthy marketplace can be created for a certain breed of Generative AI applications that will solve some of the core issues, such as: fairness, bias, safety, and security thus creating Generative AI applications that are more ethical. The combination of an open source (and therefore ethical) Generative AI model and infrastructure along with an ethical marketplace seems like a very interesting path to improve Generative AI ethics.

Conclusion

Open source is one of the leading go-to-market paths that should be evaluated for Generative AI models and applications. Choosing an open-source go-to-market strategy can support a more ethical breed of Generative AI, with unparalleled transparency.

Image generated using DALL-E 2
Prompt: business models in technology, graphic style with no text

At A Glance

key takeaways from this article

- There are multiple proven paths to monetization, each which bring differences in ethics, IP protection and market viability
- Generative AI brings a new challenge to open source: the significant initial resource investment requires a more defensible approach
- Experimentation is key: start with something that fits your intitial prospects and users. Measure what works, and adapt fast. For example, think of which features are not in the core (For Open Core), which support options exist and their prices

The Rise of Generative AI: **What Enterprise Investors Need to Know**

Written by Anik Bose & Yash Hemaraj

Image generated using DALL-E 2
Prompt: a conversation between diverse investors discussing a potential investment, water color

The release of ChatGPT by OpenAI has catalyzed a "Generative AI" storm in the tech industry. This includes:

- Microsoft's investment in OpenAI which has made headlines as a potential challenger to Google's monopoly in search.
- The recent re-release of Microsoft's AI-boosted search engine Bing to one million users which has also raised alarms around misinformation.
- VCs have increased investment in Generative AI by 425% since 2020 to $2.1bn.
- "Tech-Twitter" has blown up and even mainstream media, whose job is under the biggest threat from such advancement, carried articles around the topic.

As investors at BGV, Human-centric AI has been the foundational core of our investment thesis around Enterprise 4.0. We have dug deeply into the challenges of building B2B AI businesses in our portfolio, as well as how disruptive venture-scale businesses can be built around Generative AI.

Recently we put ChatGPT to the test on the topic of human-centric AI by asking two simple questions: What are the promises and perils of Human AI? And what are important innovations in Ethical AI? We then contrasted ChatGPT's responses with what we have learned from subject matter experts on the same set of questions (1). This analysis combined with BGV investment thesis work led us to address four important questions on the topic Generative AI.

We make the case that the timing for Generative AI is now, but Chat GPT will not replace search engines overnight. We have a strong conviction that it will unleash tremendous startup innovation that will drive enterprise productivity and GDP growth. However realizing its promise will require strong guardrails to address trust and ethical AI concerns.

Why Now?

The confluence of the sharp decline in AI hardware and software costs combined with the maturation of key generative AI technologies leads us to believe that the time for innovation has already begun. Economic downturns are periods of rich creativity and innovation; research shows that over half of Fortune 500 companies were created in downturns. We believe that this period of innovation will be fueled by a wave of AI-led enterprises required to write the next chapter of Digital Transformation.

With rapid progress in transformer and diffusion models, we now have models that are trained on billions or even trillions of models. These models can extract connections and relationships between varied sources of content, enabling AI models to generate text and realistic images and videos that a human brain may find hard to distinguish from reality. These systems' ability to generalize as well as preserve details provides an order of magnitude improvement over the "search-query and show relevant links" model.

Replacing Search Engines

The release of ChatGPT is an illustration that the timing for Generative AI is now. However while ChatGPT represents a tremendous area of innovation, it will not replace Google search engine overnight for a few reasons:

(1) Erik Brynjolffson in Daedelus (2022) titled "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence" on the promises and perils of AI and Abhinav Ragunathan, who published a market map of Human Centric Ethical AI startups (EAIDB) in the EAIGG annual report (2022)

- The answers are generic, lack depth and are sometimes wrong. Before trusting ChatGPT responses implicitly, users will need to confirm the integrity and veracity of the information sources. Already, StackOverflow has banned ChatGPT, saying: “the average rate of getting correct answers from ChatGPT is too low, the posting of answers created by ChatGPT is substantially harmful to our site and to users who are looking for correct answers.” Given the effort required to verify responses, ChatGPT’s chatbot is not prepared to penetrate enterprise (B2B) use cases.

- Putting aside accuracy, there is also the question of sustainable business models. While ChatGPT is free today and costs per billion inferences are falling sharply, running GPUs is expensive, so profitably competing at scale with Google search engine is difficult.

- Google will not stand still; they have already launched Bard with machine and imitation learning to “outscore” ChatGPT on conversational services. We’ve only seen the opening act in a much larger showdown.

- The recent release of Microsoft’s AI powered search engine has exposed key underlying issues in using ChatGPT as a search engine.

At first glance, the results look impressive, but they often lack the depth that you gain while talking to a subject matter expert.

Emergence of Ethical Issues and Human in the Loop

The accuracy concerns associated with Generative AI raise an important question - what guardrails are needed to 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 them entirely in B2B use cases due to lack of domain-specific contextual knowledge and the need for trust and verification of underlying data sets.

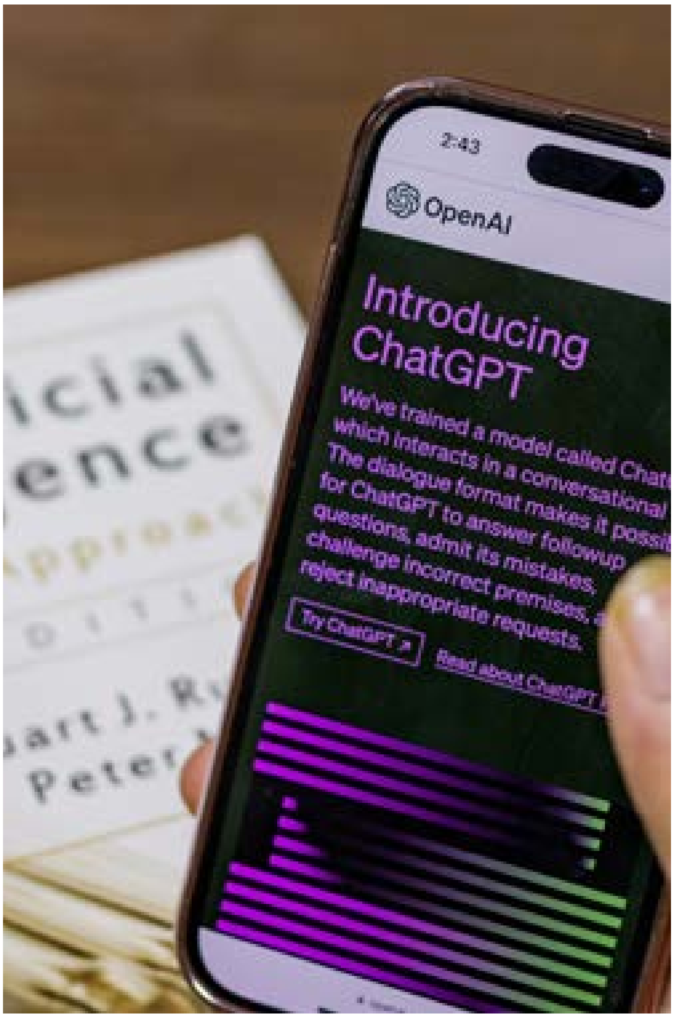
Broader adoption of ChatGPT will spur an increased demand for authenticated, verifiable data. This will advance data integrity and verification solutions, alongside a number of other ethical AI issues such as privacy, fairness, and governance innovations.

The surge in Generative AI interest will quickly prompt demands to **prioritize human values, ethics, and guardrails**. Early indications of this are:

- The recent publication from Kathy Baxter and Paula Goldman from Salesforce “Generative AI: 5 Guidelines for Responsible Development” to usher in the development of trusted generative AI at Salesforce.

- Jesus Mantas from IBM also summarizes the underlying ethical issues in his article AI Ethics Human in the Loop: “…we need a broader systemic view … that considers the interconnection of humans and technology, and the behavior of such systems in the broadest sense. It’s not enough to make parts of a decision system fair … if we then leave a risk of manipulation in how humans interact with and use those algorithms’ outcomes to make decisions.”

- Sam Altman Founder of OpenAI: “ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness. It’s a mistake to be relying on it for anything important right now. It’s a preview of progress; we have lots of work to do on robustness and truthfulness.”



Attractive investment areas for startups

The innovation prompted by the Generative AI boom and the need for trust— is poised to follow a curve similar to "ethics predecessors" like the tremendous innovation in cybersecurity in the late-2000s and privacy in the late-2010s. Analogous to Cybersecurity, Generative AI innovation will unleash startup innovation that will drive productivity gains in enterprises beyond chatbots to cover use cases like generating written and visual content, writing code (automation scripts) debugging, and managing and manipulating data. However, many of the first wave of generative AI startups will fail to build profitable venture scale B2B businesses unless they explicitly address the following three core barriers:

- Inherent trust and verification issues associated with generative AI
- Lack of defensible moats, with everyone relying on same underlying foundational models
- Lack of sustainable business models given the high costs of running generative AI infrastructure (GPUs)

It is unclear where in the stack most of the value will accrue, whether infrastructure, models, or apps. Currently, infrastructure providers (like NVIDIA) are the biggest benefactors of OpenAI.

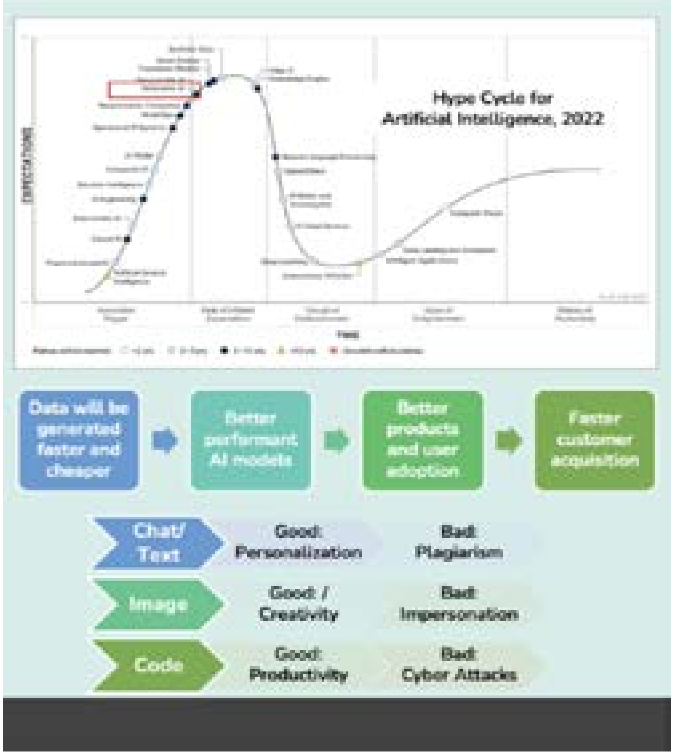
It is also unclear where startups can break the oligopoly of the infrastructure incumbents like Google, AWS, and Microsoft who touch everything, as explored in "Who Owns the Generative AI Platform?" an article published by a16z.

Successful Generative AI B2B startups mayfall into three core categories:

- **Applications** that integrate generative AI models into user-facing sticky productivity apps. Using foundation models or proprietary models as a base to build on (verticals like media, gaming, design, copywriting etc. or key enterprise functions like DevOps, marketing, customer support etc).
- **Models** to power the applications highlighted above verticalized models will be needed. Leveraging foundation models, using open-source checkpoints can yield productivity and a quicker path to monetization but may lack defensibility.
- **Infrastructure** to cost effectively run training and inference workloads for generative AI models by breaking the GPU cost curve. We will also see AI Governance solutions to address the unintended consequences of disinformation that will be created by broader adoption of tools like ChatGPT, as well as a wide range of ethical issues.

Generative AI is ushering in a novel computing model, one that turbocharges the way computers are programmed, the way applications are built, as well as the number of people that can actually put this new compute platform to work. In the case of workstations, the number of people who could put this computing model to work was measured in hundreds of thousands of people. For PCs it was measured in hundreds of millions. For mobile devices it was billions of people. The number of applications grew exponentially with each subsequent computing model.

With Large Language Models the number of productivity applications is going to grow exponentially because it will enable anyone and everyone to write their own applications. Most successful startups will be in the application layer. Especially those startups making use of the democratization of generative AI, but which take it to everyday workflows by using intelligent workflow automation and leveraging proprietary verticalized data sets to provide the most productivity improvements to end users.


Hype Cycle for
Artificial Intelligence, 2022
Data will be generated faster and cheaper
Better performant AI models
Better products and user adoption
Faster customer acquisition
Chat/ Text
Good: Personalization
Bad: Plagiarism
Image
Good: / Creativity
Bad: Impersonation
Code
Good: Productivity
Bad: Cyber Attacks

There will also be opportunities for startups to innovate at the hardware layer – to break the GPU cost curve, though these are likely to be more capital-intensive investments. Along with the opportunities for value creation there are several downside risks associated with Generative AI including:

- **Text**: while writing marketing text may be a force multiplier for martech solutions, potential plagiarism of original content may expose enterprises to liability.
- **Images**: while Generative AI can create marketing images at scale without needing product costs, the same models could create threats like human impersonation.
- **Code**: while regular snippets of code can be written automatically, the same code generation could be used to exploit vulnerabilities within enterprises.

Conclusion

Generative AI is poised to unleash a tremendous wave of innovation in human productivity in use cases like writing code, creating content, debugging, and managing/manipulating data. However, it is important to not get caught up in the generative AI hype and properly assess potential investments for both the technology and the ethics of the application.

Image generated using DALL-E 2
Prompt: a conversation between investors discussing a potential investment, water color

At A Glance

key takeaways from this article

- Generative AI is unlikely to replace Google search overnight or displace human subject matter experts. Augmenting this technology with domain expertise will be important for successful broad adoption.
- Startup innovation will be essential - Large enterprises will find it difficult to pull up an API and start using Generative AI in enterprise contexts without relying on startup technology innovation at the application layer for productivity, use cases that augment human productivity via intelligent workflow automation and proprietary verticalized data sets.
- Need for guardrails - Broader adoption of Generative AI will spur increased demand for guardrails and innovation. The tech stack of the future will address not only productivity issues, but ethical issues like privacy, fairness, and governance.

Navigating Risks and Rewards

An Intro to Using Generative AI for Data Fabrication

Written by Josh Fourie

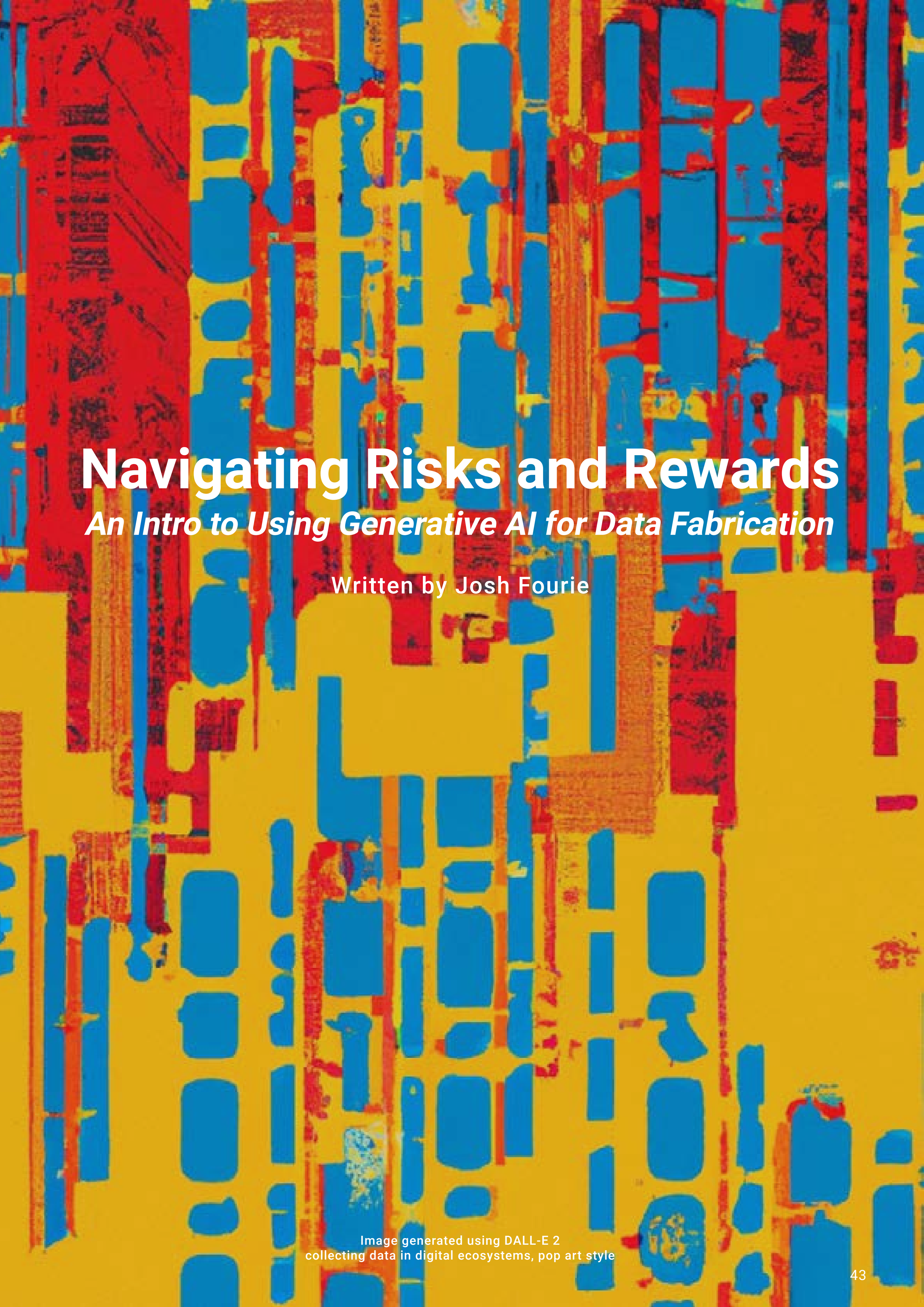


Image generated using DALL-E 2
collecting data in digital ecosystems, pop art style

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 - 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.

Breaking through the Simulation Limits of Reinforcement Learning

One interesting application of GenAI is as a tool for enriching the fabrication of simulated environments in which more sophisticated AI 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 happen 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.

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

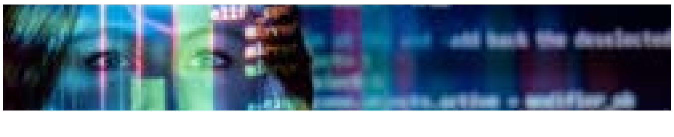
Systems like these offer a tempting opportunity for well-resourced market actors to boost the quality of their simulations to develop complex, dynamic and reactive AI systems. When thinking about or working with these kinds of systems at a high-level, I recommend taking on one broad philosophy and undertaking 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, a 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 GenAI 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 AI risks.

Over the next few years, we are likely to see systems like these gain in popularity as GenAI grows increasingly impressive in graphical fidelity, the use of synthetic data is normalized and we look to use AI 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 AI.

Image generated using DALL-E 2
Prompt: collecting data in digital ecosystems, pop art style

At A Glance

key takeaways from this article

- Generative AI has the potential to break through key constraints in simulation-based Reinforcement Learning that can unlock the next-generation of AI - but not without risks.

- We need to look behind the tools and beyond the data to understand how risk can be created and transformed through unpredictable interactions between dueling AI systems.

- To build bolder, more effective and aligned AI, it is critical that we carve out time during development for effective and continuous testing that defines the limitations of our work so that we can understand, defend and champion our risk posture.

Are you afraid of generative AI?

We can fix that.

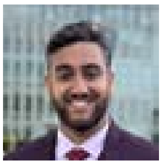
To find out how, email us at info@ethicalintelligence.co

Thank you to the humans that made this edition of the EQUATION a reality.

Helena Ward is the lead editor and coordinator for the EQUATION.

Matthew Siegel is the guest editor from EAIGG for the fifth issue of the EQUATION.

EAIGG is the supporting partner of the fifth issue of the EQUATION.

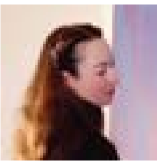


Divyansh Agarwal is a Senior Research Engineer with the interactive AI team at Salesforce Research.

Jelmer van der Linde is a Research Software Engineer at the University of Edinburgh.

Geoffrey M Schaefer heads the AI Ethics and Safety Team at Booz Allen Hamilton.

Thibaut D'Hulst is an intellectual property and data protection lawyer for Van Bael & Bellis.



Dr. Noël C. Baker is a climate scientist for the Royal Belgian Institute for Space Aeronomy, and an artist.

Ole Haaland is a robotics engineer for Prime Vision, with a special interest in AI tooling.

Alexandra Crew is an ethicist specializing in the intersection of healthcare and technology.

Matthew Douglas specialize in helping AI-driven organizations develop systematic sales strategies & processes.

Yaron Zakai-Or is the Vice President of Deepchecks, an open-source package for Machine Learning validation.

Anik Bose is a General Partner at Benhamou Global Ventures and Executive Director of EAIGG.

Yash Hemaraj is a Partner at Benhamou Global Ventures and startup mentor.

Josh Fourie is the Chief Technical Officer of Decoded.AI.

THANK YOU FOR READING

Subscribe for future issues delivered
straight to your inbox

Let us know what you think of the EQUATION

Twitter | **@ethicalai_co**
LinkedIn | **Ethical Intelligence**
Email | **info@ethicalintelligence.co**
Website | **www.ethicalintelligence.co**