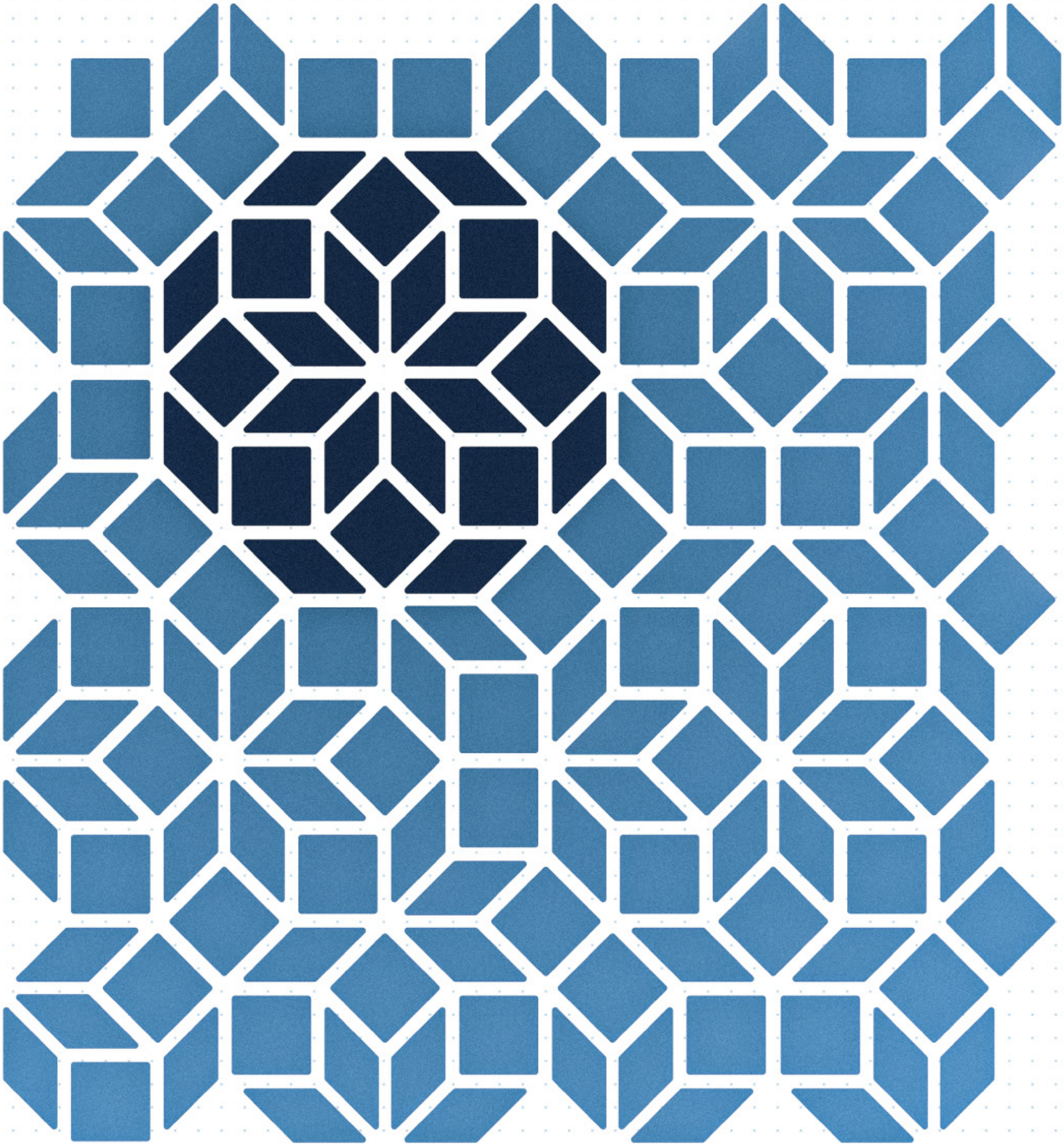
BRILLIANT MACHINES

A Pragmatic Approach to Responsible AI

WHITE PAPER



FATHOM5

CONTENTS

"When you invent the ship, you also invent the shipwreck; when you invent the plane you also invent the plane crash; and when you invent electricity, you invent electrocution... Every technology carries its own negativity, which is invented at the same time as technical progress."

– Paul Virilio

INTRODUCTION

How do we reap the benefits of new technology while minimizing its potential to do harm? This modern challenge is framed perfectly by Paul Virilio's quote above, and there is perhaps no technology that highlights the relevance of this challenge as much as AI. The last two decades have seen an explosion in the capabilities of AI, enabling new systems that can augment or replace human intelligence in a variety of tasks.[1] These kinds of systems have the potential to boost productivity, grow the economy, accelerate scientific discovery, increase efficiency, and reduce humanity's environmental footprint. Many in the field believe that AI will be one of the most disruptive technologies in human history, with Google CEO Sundar Pichai going so far as to say that "AI is one of the most important things that humanity is working on. It's more profound than... electricity or fire" [2]. Indeed, the reach of AI has quickly expanded into nearly every facet of modern life. AI helps us unlock our phones using facial recognition, browse content on the internet, find products to buy, determine the optimal route on our commute, select the best stocks to own, and even identify potential romantic partners. Increasingly, AI is being used to make highly consequential decisions such as sentencing criminals, issuing loans, and diagnosing illnesses.

At the same time, ethical and safety concerns surrounding the deployment of AI into nearly every element of modern society are beginning to mount. Many are beginning to recognize the pervasive influence these technologies have on us both as individuals and as a collective, yet there is a distinct lack of regulation or even ethical consensus surrounding AI and its acceptable applications. In fact, a growing group of philosophers, politicians, technologists, and scientists view the continued development of AI without appropriate societal safeguards as an existential risk on par with nuclear warfare and climate change.[2] Yet, one need not look to the future to think about ways in which AI might harm humanity—we are already beginning to see the effects of unrestrained development and application of these technologies. AI-enabled technologies have been found responsible for negative mental health outcomes, changes in brain structure and development (especially in children and adolescents), degraded attention and cognition, decreased social and political cohesion, widespread dissemination of false information, and the amplification of bias and discrimination. Together, these issues coalesce to threaten some of our most fundamental social and political institutions. If we continue to follow our current approach to the development and application of AI within the existing sociotechnical context, we should expect this list of harms to continue to grow in both breadth and severity as these technologies mature and become more capable and tightly integrated into our society.

What is Artificial Intelligence?

Artificial Intelligence (AI) is an extremely amorphous term in common usage—it can mean something entirely different to marketers, scientists, executives, and lay people. For the sake of this paper, we define AI as the collection of digital technologies capable of performing tasks traditionally thought to require human intelligence, such as image classification and autonomous control. These technologies are not explicitly told how to perform a specific task. Rather, they are designed with a set of principles that guide their performance within a specific domain across variations of that task, thereby exhibiting intelligent behavior.

Recognizing that AI is still in its infancy, we must seek to develop and apply AI in contexts where we can experiment with and learn about these technologies without the risk of fundamentally damaging individuals and societies at scale. Therefore, we must actively work to discover industries and applications where we can realize AI's benefits while presenting minimal potential for widespread harm of the kind described above. This paper lays out Fathom5's claim that AI in the context of industrial optimization will help solve the critical challenge of supporting a growing population in an ecologically sustainable way without the risk of widespread individual or societal harm.

> **"We must actively work to discover industries and applications where we can realize the benefits of AI while also presenting minimal potential for widespread harm."**

[1] Readers interested in learning more about the potentially revolutionary impacts of AI may wish to read [1]

[2] For more information on the field of existential risk, visit [3]

As a short note to the reader, despite our efforts to keep this paper as brief as possible, this paper is long—certainly longer than the traditional corporate white paper. However, we feel that this length is necessary to have an informed discussion that captures the nuanced challenges posed by AI. Indeed, we view finding a responsible way to harness the power of AI as the defining challenge of the 21st century, and we believe that an honest presentation of our understanding of that challenge and a realistic approach to tackling it is more important than a five-page piece of marketing material. To that end, we have tried to refrain from using hyperbole in this document and ensure that bold claims are backed by clear reasoning and research. We encourage those that are skeptical of our claims to investigate the cited resources for themselves and see if they reach the same conclusions we have. Finally, we invite readers with any thoughts or feedback to connect with us at hello@fathom5.co.

WHAT MAKES AI RISKY?

In the past decade, AI has recorded landmark performances in several areas traditionally considered to be particularly difficult for the field: Microsoft's ResNet achieved superhuman performance in the ImageNet Large Scale Visual Recognition Challenge in 2015, DeepMind's AlphaGo beat the reigning Go world champion in 2017, and natural language processing (NLP) models began outpacing traditional NLP performance benchmarks like GLUE and SQuaD in 2019. These were challenges that had seemed insurmountable to AI researchers only a few years prior, and the technologies that drive them were quickly put to use in a variety of commercial applications. At the same time, few technical or societal safeguards have been put in place to reduce the negative impacts of these technologies in the face of their rapid integration into society. Thus, we find ourselves in a position that would be analogous to having immediately given everyone a car shortly after its invention, but not having created seatbelts, airbags, anti-lock brakes, drivers' licenses, traffic laws, or highway patrol. The following sections discuss some of the most critical risks posed by AI in the current sociotechnical context.

Statistical Nature

At their core, the deep learning and reinforcement learning methods that currently dominate the field of AI are highly adaptable, statistically driven function approximation machines [4]. These models can achieve superhuman performance on a variety of tasks by ingesting vast amounts of data and iteratively adjusting millions or billions of parameters[3] to fit a function that produces the desired output given various inputs, be it the classification of an image with a correct label or the winning placement of a piece in Go. To put the size of these models and the amount of data they are trained on into perspective, OpenAI's state-of-the-art NLP model GPT-3 is composed of ~175 billion parameters and was trained on ~45 terabytes of text, including over 400 billion text fragments [5]. Though in principle somewhat analogous to the operation of biological neural networks, in practice the reasoning of statistically driven AI models diverges significantly from human cognition, which operates in terms of abstract thought and conceptual reasoning. As a result, despite achieving "superhuman" performance in many tasks, modern AI does not function like human intelligence. This has negative consequences that make its application in certain domains risky.

One consequence of the statistical nature of modern AI is that the performance of any given AI system is confined to the narrow range of tasks for which it was designed to perform (often phrased as "narrow AI") and to the statistical characteristics of the data used to train the system. In practice, these systems lack the intuitions, contextual cues, and common sense possessed by humans, embodying the phrase "common sense isn't so common." As a concrete example, image classification algorithms based on deep neural networks can be easily tricked into making comically bad classification predictions by adding random noise to an image that would be imperceptible to a human (see Figure 1). This is because deep neural networks used for image classification are optimized against massive datasets to pick up on statistical characteristics shared within image classes but unique between them. A small change in the statistical distribution of an image significantly changes the content of the image for a statistically driven function approximation machine but not for the visual cognition processes of a human. This is the driving force behind modern AI's reliance upon large quantities of data to achieve sufficient performance. As a general rule, the larger the number of parameters in the model, the larger the number of samples needed to fit the model accurately. Moreover, the examples within the training data must have similar statistical characteristics to the examples on which the model will be expected to perform in the real world. If an image classification model is trained to classify images of humans using only pictures of humans standing up, it will struggle to perform well on images of humans sitting down because the statistical

[3] A parameter is an adjustable element of an algorithm that is tuned to maximize performance. One might think of them as the "knobs" dictating the behavior of the algorithm that are carefully adjusted until the algorithm achieves the desired performance.

distribution of those images is different. Whereas you can see an object once and are likely to recognize that same object again, even under different conditions, AI models need to see the object thousands of times under every conceivable condition to achieve that same level of performance. As a result, the learning process and failure modes associated with modern AI models differ significantly from those associated with human intelligence, even for the same task.

The statistical underpinnings of modern AI also make it difficult to understand how these models make the decisions they do from the perspective of human cognition. Despite increased research into "explainable" AI in recent years, most AI models remain difficult to interpret even for those who built them, commonly referred to as the "black box" problem. The black box problem is often cited as stemming from the size and complexity of most modern AI systems. For models with billions of parameters, it is extremely difficult to understand the role played by each of those parameters and the relationships between them in shaping the model's learning and output. However, the more fundamental problem at play here beyond the sheer complexity of the model is the disconnect between purely statistical inference and conceptual understanding. To illustrate this point, consider the relatively simple case of trying to understand the moves of a model trained to play Chess. Describing why they made a certain move, a human player would describe their move in terms of the game mechanics and concepts, saying something like, "I chose to make move X because it traps their knight and gives me control of the center of the board." However, an equivalent description of an AI's move would be something like, "after observing hundreds of thousands of games to learn a probability function for the best moves, move X was found to be associated with the highest probability of a win given the current state of the board." The AI may point to the correct move, but it is up to humans to interpret the relative strengths and weaknesses of the move within the context of the game. This disconnect between statistical and conceptual reasoning makes it difficult to understand how modern AI models are solving problems in a particular context in a way that can be meaningfully understood by human beings.

As we will discuss at length later in the paper, this lack of intuition regarding how AI models work and how they can fail can make the use of AI particularly risky for certain applications. Without the ability to probe a model and meaningfully understand how and why an AI makes decisions, it is difficult to trust that the reasoning behind the model's output is functionally relevant and ethically acceptable. Moreover, it also makes it difficult to predict and prepare for circumstances in which the model might perform in unexpected ways, as in the case of Figure 1. This, too, can erode trust in a model—even if a model performs extremely well across many examples during development and testing, it may catastrophically fail in unpredictable ways when deployed in the real world.

Lack of Regulation

Governments worldwide have struggled to keep up with AI's rapid innovation and adoption. To date, no major government has enacted a comprehensive regulatory framework specifically addressing AI concerns. Instead, most rely on the extension of existing regulatory regimes to AI. This is problematic because existing regulations are often difficult to apply in the context of AI as they lack consideration for factors specific to AI. For example, the

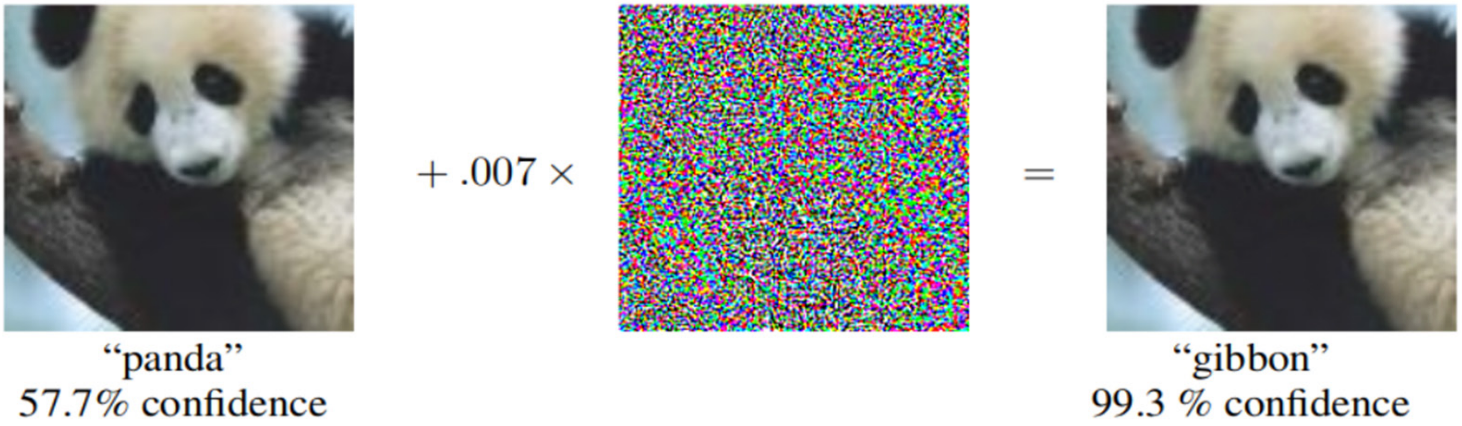


Figure 1: Slightly altering the statistical distribution of an image can cause major issues for modern image classification methods. Here, an image classification model correctly classifies the original image as a panda. However, slightly changing the image by adding a small amount of noise causes the model to misclassify the panda as a gibbon with high confidence, despite the changed image appearing identical to the original for a human viewer. (image credit [6])

United States Equal Credit Opportunity Act (ECOA)—introduced in 1974—prohibits "credit discrimination on the basis of race, color, religion, national origin sex, marital status, age, or because a person receives public assistance" [7]. Therefore, it is illegal to directly factor in so-called "sensitive characteristics" like race when making lending decisions. However, it is possible for AI models that take thousands of non-sensitive data points as inputs to indirectly infer sensitive characteristics with a high degree of accuracy. Data related to an individual's spending habits, social network, and browsing history can allow a model to infer race, gender, marital status, etc., and the difficulty understanding these models' decision-making process (as discussed above) can make it extremely difficult if not impossible to determine whether such inferences are being made inside a model and, if they are, how these inferred characteristics are affecting the model's outputs.[4]

The pragmatic challenge of developing effective policy surrounding AI is exacerbated by disagreement over ethical questions that arise from the application of these technologies—especially at the international level. Who can use AI, for what purposes, and with what characteristics? General agreement on the answers to such normative questions is a necessary backdrop for the formulation of policy that guides AI in a direction that we as humans agree is desirable. Over the past few years, many organizations across the private, public, non-profit, and academic sectors have begun to release guidelines for the ethical use of AI that are meant to address these questions. However, a study surveying global AI ethics guidelines in 2019 found that despite emerging consensus around a core set of eleven general ethical principles, there are "substantive divergences… in relation to four major factors" [9]. These factors were 1) how ethical principles are interpreted, 2) why they are important, 3) what issue, domain, or actors they pertain to, and 4) how they should be implemented. For example, the majority of the ethical guidelines included justice as a fundamental principle of ethical AI yet varied greatly in their definitions of what "justice" means in the context of AI and how it can be ensured. The difficulty associated with formalizing high-level moral principles into a more granular understanding of which actions are and are not acceptable under specific circumstances is nothing new—it has been a source of moral and political debate for thousands of years.[5] Yet, these questions gain a renewed importance and urgency in the face of a technology capable of rapidly replacing or augmenting human cognition to make decisions of direct or indirect moral consequence.

Recent years have seen the first steps toward regulations that specifically consider factors relevant to AI. Some of these regulations, such as the EU's General Data Protection Regulation (GDPR) and the proposed Digital Services act, are limited to the narrow issues of data rights and content recommendation algorithms, respectively. Others are aimed at creating comprehensive regulatory frameworks for AI technologies. In 2021, the European Commission proposed the Artificial Intelligence Act to provide the EU with a legal framework for AI, the Brazilian Congress passed a similar bill (with the Senate set to vote sometime in 2022), the Cyberspace Administration of China released a three-year road map for governing all internet algorithms, and AI-specific laws have begun to emerge on a local and state level in the United States and elsewhere. However, given AI's global reach and implications, truly effective AI regulations will require a high level of international coordination and cooperation rather than a patchwork of local, state, and national laws [11].

If the historical timeline for progress on nuclear arms reduction work is any indicator, it will take years to form international consensus on AI regulation and longer still for these regulations to be effectively implemented, enforced, and refined, especially in the context of our relative lack of understanding with regard to AI technologies. Moreover, the rapid pace of AI development presents a more general challenge for a regulatory paradigm based on hundred-page pen-and-paper documents and annual legislative sessions. As a result, we are far from a truly effective and comprehensive regulatory paradigm for AI. Given the central role played by government in mitigating the negative effects of technology, the relative lack of effective AI regulation for the foreseeable future significantly increases the risk that harmful applications of AI will remain unaddressed.

Bad Incentives

When acting within complex systems, human beings reliably follow incentives as they perceive them—this is one of the fundamental insights of behavioral economics. For companies, these incentives take the form of profit over various timescales. The individuals within a company are not only likely to make decisions that benefit shareholders through improved profit, they have a fiduciary duty to do so. This is important to understand given that the commercial sector currently leads the field of AI in both development

[4] For an in depth look at the ethical and legal issues surrounding the use of AI for credit scoring, see [8]

[5] See [10] for further reading on this point and its relevance to AI ethics

and application, investing more than academia and government spending combined. In 2020, the United States commercial industry spent over $80 billion on AI, while non-defense[6] spending by the federal government was only $1.5 billion that same year [12]. Moreover, the private sector's lead in developing and deploying AI is increasing. In 2019 65% of AI PhD graduates in North America went into industry, up from 44.4% in 2010 [13]. Though we certainly want companies to behave ethically, we must also be realistic in understanding that profit incentives sometimes contradict ethical behavior, as was exemplified by Google's controversial firing of leading ethics researchers Timnit Gebru and Meg Mitchel and the ensuing turmoil surrounding its ethical AI team [14]. Therefore, to ensure that AI is developed and deployed responsibly for the benefit of humanity, it is critical that the profit incentives driving the private sector are aligned with this outcome.

Unfortunately, there are many markets for which this is not the case. One obvious example is markets that operate within the so-called "attention economy." In the attention economy, companies make revenue by harvesting the attention of users and monetizing the valuable data it generates. For example, the vast majority of Facebook's revenue—about 97.9 percent in 2020 [15]—comes from ads, and a given advertiser is willing to pay more or less depending on how much attention they believe a given ad is likely to capture from their target audience. The more of your time and attention Facebook can capture, the more it knows about you from the data you generate when using the platform, and the more money it will make. Indeed, social media platforms' real customers are the advertisers, not the end user, because the advertisers are actually paying for the service. Notice that this is different from more traditional business models wherein the user is the one who pays for the service, which in turn incentivizes the company to maximize the benefit of the service for the user. "If it's free, you are the product," as the saying goes. As will be discussed in the following sections, the skewed incentives that arise from this business model have led social media platforms to deploy AI as part of a global-scale effort to maximize engagement and attention in ways that are often detrimental for both individual users and society as a whole.

It is important to acknowledge that the attention economy is not inherently bad, nor is it the only market force that can lead private companies to take actions that harm the general public. However, it is a well-studied case that serves to highlight the dangers of allowing maligned business models and incentives to drive the development of AI, especially in the absence of coherent regulatory and ethical frameworks. If business incentives stemming from markets that are maligned to the public good are left to drive the way AI is applied, then AI will often be applied in ways that are maligned to the public good.

Scalability

As software-based digital technologies, AI systems can be rapidly scaled and disseminated in ways that simply aren't possible for physical systems.[7] Consider the fact that Facebook had almost 2 billion active daily users in the first quarter of 2022 [17]. At the press of a button, Facebook can release an update to its content recommendation algorithm that affects nearly 1/4th of humanity within a single day. This means that the consequences of an algorithm—both good and bad—can almost immediately reach global significance. At these scales, an algorithm's negative effects can be particularly insidious—even a small change in the behavior of millions or billions of individuals can have a huge effect in the aggregate. Innovation in AI disseminates far more easily than physical technologies as well. For example, we have known how to generate power through splitting atoms for a long time, but to take advantage of this knowledge by building a nuclear power plant (or a nuclear weapon) still requires massive amounts of capital, labor, coordination, and technical sophistication. By contrast, one can utilize many cutting-edge models in AI with nothing but a laptop and some AWS credits.[8] Moreover, aided by modern hyperscale computing infrastructure, AI models can be deployed at scale by small teams of engineers with little oversight or public accountability.

> **"At the press of a button, Facebook can release an update to its content recommendation algorithm that effects nearly one fourth of humanity within a single day."**

[6] Data on defense spending related to AI is not publicly available

[7] For further reading on this point, see [16]

[8] This is especially true using a technique known as transfer learning. Transfer learning takes a pretrained state of the art model such as Google's MobileNet V2 and retrains only a small portion of the model to tune its performance to a new task. To see for yourself how easy this really is for someone with a basic foundation in programming, see this Google tutorial: https://www.tensorflow.org/tutorials/images/transfer_learning

The scalability of AI really is the critical point that makes the other characteristics outlined above so concerning. The process of utilizing AI for more and more applications is really the process of integrating novel decision-making agents into our society. Cumulatively, the risks outlined above aggregate to create a world in which this process is taking place in an unregulated environment, driven by potentially maligned incentive structures, with relatively poorly understood agents that display an intelligence very different from our own—and because these technologies can be scaled so quickly, billions of people are being exposed to the risks associated with AI in a relative blink of an eye.

HOW ARE THESE RISKS REALIZED AS HARMS?

Potential risks like the ones outlined above are cause for concern and evaluation. But when these risks are realized as measurable harms, it is time for action. The following sections outline only a few ways that deploying AI in the current sociotechnical context is causing significant individual and societal harm. What should be clear from the following sections is that these harms—as bad as they are today—are only the beginning of what is to come. Unless we make significant changes to how we handle AI application and development, we will almost certainly see these harms amplified as AI technologies continue to advance and become increasingly widespread.

Psychological Health

A growing body of evidence[9] shows that excessive usage of digital technology has negative impacts on psychological health, manifesting as negative mental health outcomes, degraded cognitive abilities, and even structural changes in the brain. These effects are especially pronounced in children and adolescents whose brains are undergoing critical periods of development that will affect them for the rest of their lives. Indeed, studies have found that preschool-aged children with higher levels of screen time showed delays across key developmental measures, including language, problem-solving, and social interaction, along with brain scans showing physical changes in areas of the brain associated with language in these populations [19], [20]. Social media in particular is a driver of these problems. Both academic [21] and internal company [22] research has shown that Instagram use is linked with significantly higher rates of eating disorders in teenage girls. Moreover, a randomized experiment on over 1,600 American adults who used Facebook for up to an hour a day found that one month away from the platform led to a significant improvement in emotional well-being and a large, persistent reduction in post-experiment Facebook use [23].

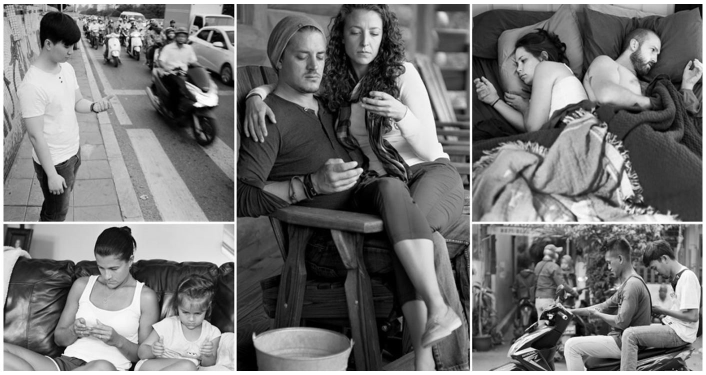


Figure 2: Photography collection by Eric Pickersgill wherein phones and tablets have been photoshopped out of the images to highlight the absurd amount of time and attention captured by these devices. (image credit [24])

Of course, "digital technology" is not the same as AI, and many of the negative effects associated with the use of digital technology are mediated by factors other than AI. For example, much of the harm caused by social media use has to do with its ability to expose an individual to millions of other people leading to unhealthy levels of social comparison for which we are not adapted. However, the AI-powered content recommendation algorithms that underpin these sites massively amplify this harm by exploiting thousands of hours of personalized behavioral data to stoke social comparison and maximize our time on site as part of the attention economy. In doing so, these algorithms hijack the limbic system in our brain in ways that are associated with the structural changes [25] and behaviors [26] characteristic of addiction. If social comparison is the drug, then your personalized content recommendation algorithm is the dealer whose goal is to get you to buy more. It is no wonder that Chamath Palihapitiya, a former vice president of user growth at Facebook, said: "The short-term, dopamine-driven feedback loops that we have created... [are] eroding the core foundations of how people behave by and between each other. I can control my decisions, which is why I don't use that sh*t. I can control my kids' decisions, which is that they're not allowed to use that sh*t" [27].

[9] For a more comprehensive overview of these harms, see [18]

Manipulation

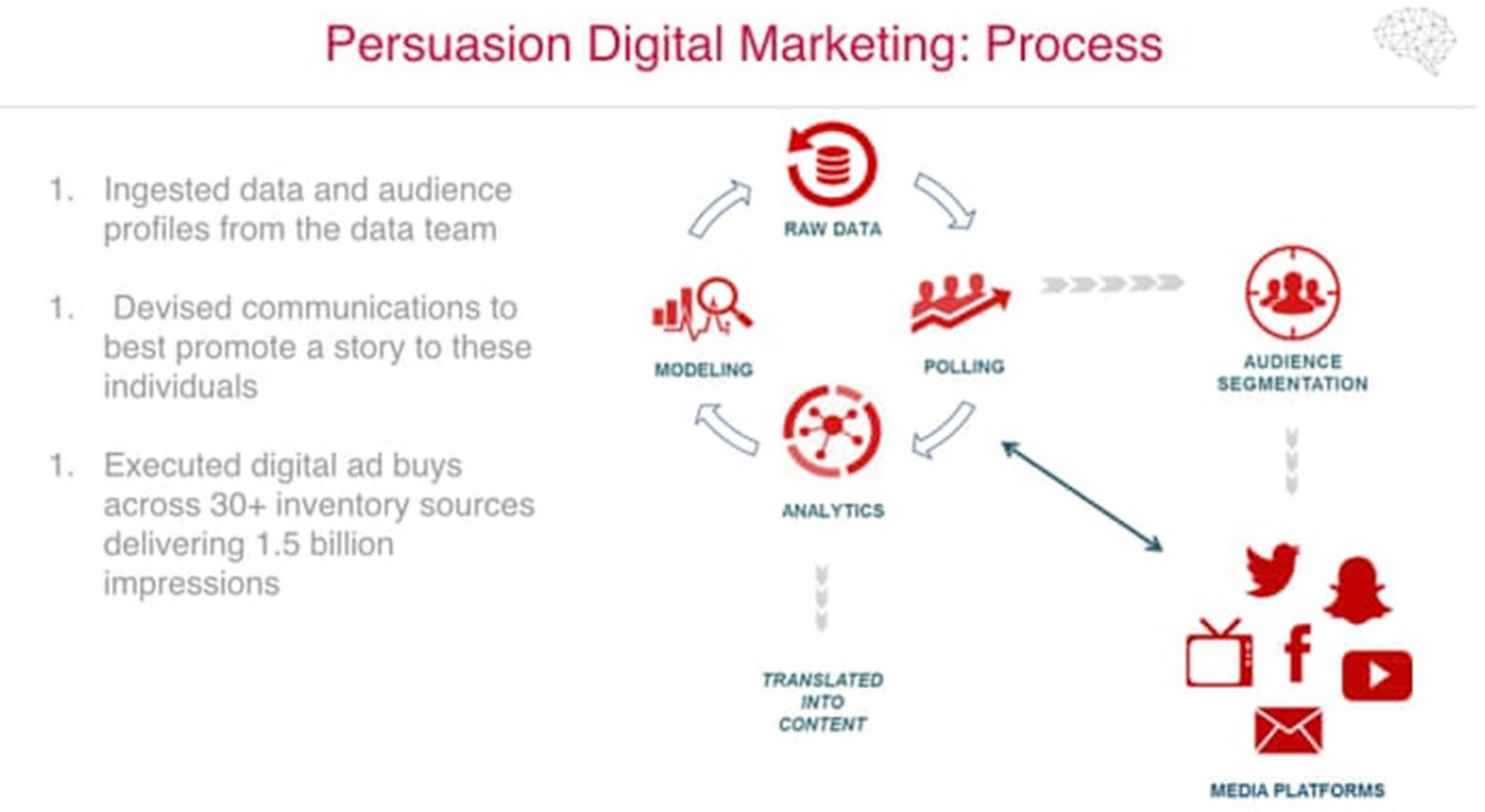
There is growing concern that algorithms and the vast amount of data they have access to can be used to alter people's opinions and actions in a way that degrades their individual autonomy. Of course, some amount of persuasion has been an acceptable part of society for a long time. After all, what is an ad if not an attempt to persuade you to hold a given opinion or buy a specific product? Yet, there are differences both in degree and in kind between traditional advertising strategies and those driven by AI. Take, for example, an ad placed on a traditional cable television channel for a particular brand of engagement rings that are on sale. The company that bought the ad picked that particular channel based on general information about the channel's demographics, perhaps that the channel's viewers are predominantly men in their late twenties. Some will be persuaded to buy that brand, some will be outside of the intended customer demographic, and some will have muted the TV while they go to the kitchen. Now consider that same ad delivered via your favorite social media platform using an ad placement algorithm. Your platform activity and social network connections indicate that you are a straight 28-year-old male who has been in a relationship for two years[10]; your browsing history (off the platform!) indicates that you have been shopping around for engagement rings; it is late and your usage metrics over the past hour indicate that you are getting drowsy and more susceptible to persuasion. The platform primes you with some specific content: a post by a friend who recently got married, an article about the best honeymoon spots, and another article describing coming increases in fine jewelry costs. Finally, you see the ad for engagement rings on sale—but not just any ad: this ad has been tailored to your personal preferences through A/B testing[11] of multiple ad layouts with thousands of other users. In fact, the woman in the ad looks similar to your soon-to-be fiancé, and she is pictured with the ring and a huge smile across her face. Of course, you click the ad and buy the ring. The difference in degree between this ad and the TV commercial comes in the form of precision. The AI-driven ad is able to be targeted at a highly specific audience in a highly personalized way. But the difference in kind comes from the feedback loop created by your interaction with the social media platform, which allows the platform to measure your behavioral response to content and ads in real time and adjust accordingly.

For many, the differences described above mark the philosophical line between persuasion and full-blown manipulation, and a growing body of scientific studies support this line of thought. In one study, researchers were able to train algorithms to manipulate study participants' behavior across three game-based experiments testing action selection, response inhibition, and social decision-making [29]. This is especially concerning when one considers that the manipulative tactics outlined above are also applied to individual decisions that have significant consequences for broader society, such as which candidates to vote for in an election. Indeed, this was the business of Cambridge Analytica, the now notorious political consultancy firm that utilized AI to deliver highly targeted and personalized ads across platforms like Twitter, Facebook, and Snapchat and monitored engagement in real time to adjust the ad delivery strategies in support of the candidates that hired them (see Figure 3 below) [30]. But manipulation goes beyond targeted ads: studies indicate that how we access information also strongly affects our preferences and behaviors. One study investigating the power of recommendation algorithms to influence preferences found that "viewer preference ratings are malleable and can be significantly influenced by the recommendations received" [31]. Another study found that altering the order of search engine results could influence the behavior of undecided voters by more than 20% [32]. What's worse, these search ranking biases could be masked so that participants showed no awareness of the manipulation. The ability to influence voter decisions by 20% is an incredibly powerful one, especially considering that democratic elections are often decided by a difference of only a few percentage points. Indeed, spending metrics reflect the power search engine rankings can have: In 2021, total spending on search engine optimization (SEO) in the US was estimated to be $52 billion [33], while total spending on government lobbying that same year was only $3.7 billion [34]. In other words, companies found it orders of magnitude more valuable to influence Google's search algorithm than the US government.

> **The ability to influence voter decisions by 20% is an incredibly powerful one, especially considering that democratic elections are often decided by a difference of only a few percentage points.**

[10] See [28] to understand how data generated through platform usage can be used to predict these traits

[11] A/B testing, sometimes called split testing, is technique wherein different users are exposed to different versions of some target variable, such as an advertisement or website layout. The behavior of these users is then measured for different presentations to identify the one with maximal intended effect.



Persuasion Digital Marketing: Process

1. Ingested data and audience profiles from the data team

1. Devised communications to best promote a story to these individuals

1. Executed digital ad buys across 30+ inventory sources delivering 1.5 billion impressions

RAW DATA
MODELING
POLLING
ANALYTICS
AUDIENCE SEGMENTATION
TRANSLATED INTO CONTENT
MEDIA PLATFORMS

Figure 3: Slide taken from a Cambridge Analytica PowerPoint presentation highlighting its digital strategy for supporting political candidates. (image credit [30])

Polarization

The way individuals access and share information can have a significant effect on the polarization of a society. Many are familiar with the concept of "filter bubbles," wherein personalized content recommendation algorithms continuously feed someone content that is aligned with their existing beliefs, ultimately resulting in a state of intellectual isolation. Google—the de facto knowledge lookup for settling any argument—gives each of us personalized search results, and your social media feed is based on the browsing habits of you and your own social network. Yet, the problem goes beyond stunting the formation of new opinions informed by diverse perspectives. Content recommendation algorithms have been shown to direct users to consume increasingly radical content. One study found that YouTube's content recommendation algorithm caused users to "consistently migrate from milder to more extreme content" based on an analysis of over 2 million recommendations and 72 million comments [35]. These findings match reports generated internally at Facebook in 2016, which found that "64% of all extremist group joins are due to our recommendation tools… Our recommendation systems grow the problem" [36]. Indeed, there is a general trend toward the promotion of anger and outrage across various social media platforms, with studies indicating that anger and outrage spread faster on both Twitter [37] and its Chinese equivalent Weibo [38]. Though these effects are partially driven by human psychology, especially in crowds, they are amplified by these platforms' use of AI to maximize user engagement. But don't take our word for it; another internal Facebook report said in 2018 that "Our algorithms exploit the human brain's attraction to divisiveness," which fed users "more and more divisive content in an effort to gain user attention and increase time on the platform" [36]. See Figure 4 for an illustration of just how polarized US Facebook users are.

Making matters worse is the prevalence of disinformation on these platforms, amplified by the existence of fake profiles controlled by AI, commonly referred to as "bots." By assembling large numbers of bots and deploying them as part of a coordinated effort to push false narratives or simply pollute the platforms with conflicting information, one person or a small group can increase the spread of misinformation by many orders of magnitude. Indeed, studies have shown that this kind of "information gerrymandering" can have outsized effects on collective decisions and voting patterns [39]. Russia

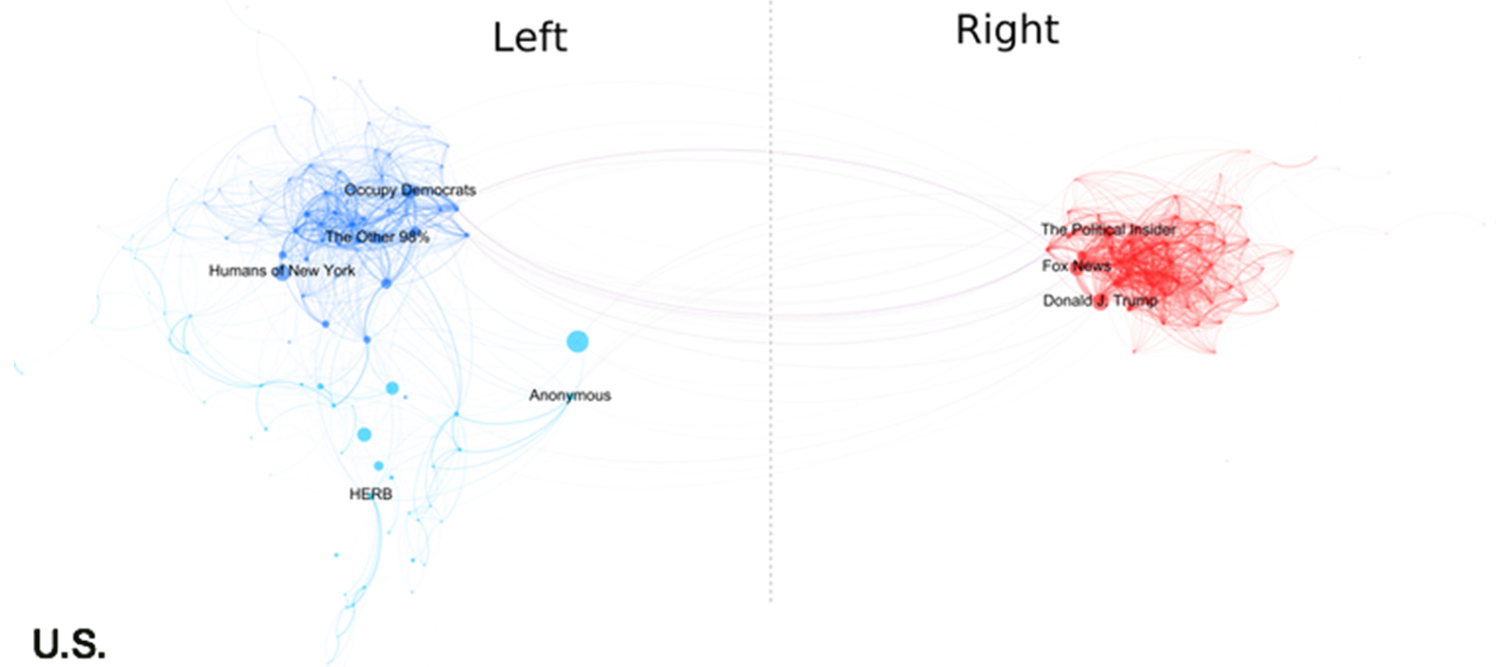


Figure 4: Graph illustrating the polarization of US Facebook users. Each node represents a Facebook page. Larger nodes are more popular pages. Two nodes are linked and grouped when users like a post from both pages, so a cluster of nodes represents a community of users interested in similar topics. (image credit [39])

used such tactics as part of a sustained disinformation campaign meant to seed division in American society surrounding the 2016 election. Between January 2015 and August 2017, 50,528 Russian bot accounts generated 3.8 million tweets related to the 2016 US presidential election, representing about 19% of the tweets related to the election during that time [40]. At the same time, Russian troll farms aided by bots generated approximately 80,000 Facebook posts reaching an estimated 126 million US citizens—more than a third of the entire US population [41]. But misinformation peddled by bots need not be part of a nation-coordinated effort to be harmful—a high prevalence of bots spreading misinformation derails our ability to reason about some of the most important issues of our time more generally. In the weeks following America's withdrawal from the Paris Climate Agreement, suspected bots accounted for roughly a quarter of all climate change-related tweets [43], and bots were found to reference low-credibility sources at much higher rates than high-credibility sources when tweeting about COVID-19 [44]. Without the ability to agree upon the facts, it is difficult to imagine how we will be able to individually and collectively reason our way through an election, a public health crisis, or an existential threat.

The combination of intellectual isolation via filter bubbles, algorithms that amplify the propagation of outrage and steer users toward increasingly extreme content, and prolific amounts of disinformation amplified by bots creates an environment on many online content platforms that is extremely conducive to negative types of polarization, reducing the possibility for productive discourse, promoting tribalism, and decreasing our ability to reason through issues. Though certainly not the only cause, it is clear that AI, as deployed by many online content platforms, has exacerbated underlying issues contributing to growing polarization.

Fairness

As AI is employed to augment or replace human cognition across an increasingly wide range of tasks, its ethical shortcomings have become glaringly apparent. In particular, racial and gender biases (among others) have been found in AI systems making decisions regarding criminal sentencing, hiring, credit scoring, healthcare spending, internet search results, and even facial recognition. Often, these biases are a consequence of modern AI's reliance on data. When the data are taken from a society that exhibits biases against certain groups, those biases are reflected in the statistical characteristics of the data and, in turn, the algorithms trained on it. For example, one study found that a healthcare algorithm used to predict which patients would require extra medical care heavily favored white patients over black patients, despite not considering race as a factor [45]. The issue stemmed from the algorithm's use of healthcare cost history to predict healthcare needs, which is predicated upon the assumption that white and black patients would spend similar amounts for similar levels of care. However, this assumption ignored racial disparities in healthcare spending, resulting in the algorithm's predictions favoring white patients. Similarly, an analysis of an algorithm called COMPAS used to predict the likelihood of a criminal defendant's likelihood of recidivism (the likelihood of committing another crime) found that the false-positive rate for black offenders was nearly twice as high as it was for white offenders (45 versus 23 percent), among other racial disparities in performance of the algorithm [46]. Systems like COMPAS are intended for use by judges to help make criminal justice decisions ranging from setting bail to approving parole and even setting sentences.

To deploy a biased algorithm to make decisions of this kind is to literally encode bias into our society. These systems perpetuate bias through a sort of self-fulfilling prophecy: biased data leads to biased algorithms, which in turn leads to biased decisions, which in turn generates more biased data. Though it is possible in principle to remedy algorithmic bias, it is extremely difficult in practice. For starters, it is often infeasible to obtain unbiased datasets to represent an unbiased "ground truth." Moreover, the opaque nature of these models discussed above stands in the way of identifying and remedying the elements of a model that contribute to its biased decision-making process. This is made even more challenging by the fact that not all bias is created equal. Some forms of bias are acceptable—even necessary—to create an accurate system. For example, we are "biased" against lower-income applicants when issuing loans because we believe income is a fair and relevant factor to consider when trying to determine the likelihood of default. So really, we are trying to shape an algorithm to be biased with regards to certain characteristics of the input data but not others. Therein lies the final and perhaps most important challenge of designing fair algorithms: who decides what is "fair"? When an algorithm is used to make ethically consequential decisions, there is an implicit judgement being made that this algorithm behaves in an ethically acceptable manner. Yet, as was discussed above, there is a lack of societal consensus regarding many of the ethical issues these algorithms bring to the forefront. In the absence of a consistent ethical or regulatory framework, the judgement of whether this algorithm behaves ethically is often left up to the team of individuals who created it. In the current societal context, this is most often a small group of engineers and product designers distinctly lacking in diversity and with no accountability to the public.[12]

The ethically loaded decisions outlined above affect peoples' lives in a fundamental way. Who has access to the highest quality healthcare, who has access to various financial opportunities, who should be imprisoned, and for how long? These decisions are at the core of our society, and entrusting those decisions to a young technology that is highly scalable and has known flaws is downright irresponsible. A biased judge is one thing, an institution that perpetuates biased judgements is another, but a biased algorithm developed by a handful of engineers and deployed into courts across America? That is unprecedented.

> **A biased judge is one thing, an institution that perpetuates biased judgements is another, but a biased algorithm developed by a handful of engineers and deployed into courts across America? That is unprecedented.**

A PRAGMATIC SOLUTION

If you have been following the paper thus far, you are probably feeling pessimistic about the prospects of AI. These technologies are clearly causing or contributing significant amounts of harm to individuals and society. The companies behind many of these harms are unlikely to make significant changes to their products and business models of their own volition because their incentives are misaligned, and we are far from a place where we can implement effective regulatory frameworks to curb negative outcomes. Moreover, these technologies present significant technical challenges for which we are unlikely to have solutions in the near future. AI has the potential to address some of the biggest challenges of the 21st century, but it might seem unclear how we can realize this potential without causing significant damage in the process.

However, it is imperative to realize that these harms are not inevitable consequences of AI. Rather, they are the result of the widespread application of a powerful but immature technology in areas that can cause grave harm. The reality is that the widespread application of AI is an unprecedented sociotechnical experiment. Indeed, AI philanthropist and former CEO of Google Eric Schmidt has said, "We are playing with the information space of humans. We are experimenting at scale without a set of principles as to what we want to do"[47]. Thus far, that experiment has been conducted directly on billions of people in ways that have pragmatic and ethical implications for the functioning of our society

[12] See chapter six of [13] for an overview of diversity in AI

and institutions. How we access information, the ways we connect and communicate with each other, decisions about how resources and opportunities should be allocated and to whom—these functions are at the core of society, and it should be unsurprising that the application of AI to these functions with the attitude of "move fast and break things" has proven to be problematic.

Trial and error in the real world are certainly a necessary element of learning how to utilize AI responsibly, but these errors must take a less costly form than false imprisonment or the erosion of democratic institutions for AI to bring a net benefit to humanity. As such, there is an urgent need to identify "sandbox"[13] industries and application areas where we can experiment with and better our understanding of AI in the real world with minimal risk of causing the devastating harms outlined above. These sandboxes will give us a safe environment to iteratively test new innovations and safeguards in AI and understand their impact in the real world before exporting those solutions to higher-risk industries and applications. Moreover, by picking sandbox industries and applications for which the use of AI has high potential to benefit humanity, the sandbox approach would allow us to realize many of the benefits of AI even as we pace the adoption of these technologies to match the development and maturity of critical sociotechnical safeguards.

THE INDUSTRIAL SECTOR AS AN IDEAL SANDBOX

Fathom5 believes that the industrial sector is an ideal sandbox industry for learning how to apply AI in a responsible manner. The potential harm that could be caused by AI in the industrial sector is well-characterized and has little chance of propagating to cause the kind of widespread harm to individuals and societies outlined above. At the same time, successful application of AI to optimize industrial operations could play a central role in tackling some of the largest challenges facing humanity in the 21st century. The following sections describe some of the key characteristics that make the industrial sector an ideal sandbox for AI development, as well as some of the risks associated with industrial AI and mitigating factors for those risks.

Objective Good

An optimized industrial sector that allows humanity to produce more goods and services essential to modern life using fewer inputs is objectively good for humanity. One might debate whether the ability of any individual to broadcast their opinion to millions of other people is ultimately good or bad or whether a credit scoring system that is overall more accurate in the aggregate justifies the risk of algorithmic bias, but few would argue against the benefit of an optimized industrial system that is able to make more with less. If you are a capitalist, you make more money. If you are an economist, you boost economic productivity. If you are an environmentalist, you reduce environmental impact. If you are a consumer, you gain access to essential goods and services at a decreased cost. If you are a humanitarian, you are able to provide clean water, food, and energy to a larger portion of humanity. In other words, the incentives of industrial optimization are aligned with the best interests of humanity as a whole. But more than being universally beneficial, industrial optimization addresses one of the crucial challenges facing humanity and the environment in the 21st century. According to the UN, the global human population is expected to reach nearly ten billion by 2050 [48]. If we try to produce enough clean water, food, energy, medicine, and other essential goods and services for this number of people using an industrial base operating with the current level of efficiency, it will kill our planet. We must find ways of producing and delivering more of these essential goods and services using fewer inputs and with a reduced environmental footprint to accommodate an increasing world population, and AI has the potential to play a central role in improving industrial efficiency.

Not only are the goals of industrial optimization objectively beneficial to humanity, but they can also be expressed in quantifiable and objective terms. This is critical because it allows us to measure the performance of an industrial AI system against a desired outcome directly rather than through a set of subjective proxy metrics. For example, one can directly quantify how successful a given AI is at maximizing the amount of food produced by a plot of land given a fixed amount of water and fertilizer. On the other hand, it would be difficult to directly measure how successful an updated content recommendation algorithm is at promoting a "healthy democracy" because the "health" of a democracy can only be measured in terms of subjectively chosen proxy metrics. Should the success of the algorithm be measured in terms of increased voter turnout? Or decreased polarization? Or maybe some weighted combination of both? If chosen wrong, the optimization of these proxy metrics can actually be counterproductive to the intended outcome of applying AI to these issues. While there is always the risk of

[13] This term is borrowed from software engineering, where a "sandbox" is an isolated development environment that engineers can use to experiment and test changes to software without affecting the functionality of the system that is actually deployed to users

unintended consequences when applying AI, the ability to define an industrial AI's objective function[14] directly in terms of desired outcomes significantly reduces the chances of deploying AI systems that are misaligned with intended outcomes.

Nature of Potential Harm

The nature and scale of the harms realized by the leading applications of AI discussed above are simply not possible as the result of AI applied to the industrial sector. It is difficult to imagine how a decision to increase the revolutions per minute of a pump could cause widespread mental health issues or alter the outcome of an election. Though the industrial sector forms an incredibly important layer of modern society, it doesn't play a central role in our elections, it doesn't capture the attention of billions of people for hours each day, it doesn't affect the way we share information and reason about the world, and it doesn't influence decisions that dictate individual opportunity, freedom, and autonomy. The problems to be solved with AI in the industrial sector are technical in nature, not social, political, or philosophical. Decisions made using machine data for machine control aren't fraught with the same ethical complications as decisions made using human data for human control. There is an obvious difference between a biased dataset that results in a batch of misplaced drill-holes and a biased dataset that results in a segment of society being systematically denied access to credit. Indeed, it is interesting to consider whether an algorithm like COMPAS would ever have been allowed in American courts had the issue of AI bias been discovered in an industrial setting before we began developing AI to assist with legal decisions.

When and where harms do occur, the combination of objective measures of performance and a beneficial incentive structure means that they are likely to be quickly detected and remediated. Industrial companies are incentivized to closely monitor every aspect of their industrial processes, and the company is highly incentivized to remedy or remove any application of AI that harms this process rather than benefitting it. An AI system that results in misplaced drill-holes, reduced efficiency, or increased emissions won't last long. In contrast, it has taken years for the subtle yet significant effects of social media platforms on mental health to be identified because social media platforms are incentivized to monitor and maximize user engagement, not user well-being. Now, even as these harms have been brought to light, these platforms have done little to remedy these harms because it is not in their interest to do so. These differences mean that the industrial sector is much less likely to present and perpetuate the insidious forms of harm associated with the current leading applications of AI.

> **The problems to be solved with AI in the industrial sector are technical in nature, not social, political, or philosophical. Decisions made using machine data for machine control aren't fraught with the same ethical complications as decisions made using human data for human control.**

Automation Precedent

Industrial control systems began to automate starting in the 1970s, and most modern industrial control systems incorporate a significant degree of automation. In these systems, humans have programmed specific automation instructions that dictate what operations the system is to perform given its current state, essentially saying, "if X, then do Y." Therefore, the application of AI to industrial automation builds upon over 50 years of experience using algorithms for machine control. The only difference between the current automation paradigm and one based on AI is that AI is able to learn from massive amounts of data generated by industrial systems to optimize these control algorithms in real time. Thus, the application of AI to the industrial sector represents a difference in degree rather than a difference in kind with respect to established technological practices. This is important because we already have a fairly good idea of the risk profile associated with industrial automation and can therefore understand and mitigate the risks AI might pose as an extension of that risk profile (see section below for further discussion of these risks). Furthermore, it means that the industrial sector already has a robust legal and regulatory framework surrounding issues related to automation, such as data ownership and liability. For example, there is a clearly defined legal framework for the determination of liability in the case of an automated industrial system that causes an accident on the factory floor. In contrast, determining liability in cases of teen harm caused by social media platforms is much trickier—how is fault distributed

[14] An objective function is used to measure and optimize the performance of an AI system. For example, an AI meant to predict equipment failures might take as input various measurements of the equipment to produce an estimate of when the equipment will fail. The objective function would capture the difference between predicted times of failure to the actual times of failure, and the AI would be adjusted accordingly to minimize the objective function (and therefore maximize the performance of the AI).

between the parents, the accounts producing harmful content, the algorithms that push that content, and the platform that hosts it? Because online content hosting platforms and the problems they can contribute to are relatively new, there is no solid legal framework in place to determine liability.

Known Risks

Of course, any application of AI comes with risks. For the industrial sector, two risks in particular stand out as being of primary concern: disruption of the labor market and the introduction of a centralized failure mode into critical industries. Fathom5 has thought deeply about these risks and believes that they are justified by the potential benefit of industrial AI both for the industrial sector specifically and for the long-term future of humanity.

Labor Market Disruption

Disruption of the labor market as a result of increasing automation driven by AI is of chief concern both within the industrial sector and as a broader macroeconomic trend. Previous waves of automation have disrupted jobs and even entire industries, but history has shown time and time again that these jobs were ultimately displaced rather than eliminated—new markets and industries quickly appeared to replace the jobs that were lost, and the labor market remained stable (or grew) in the long run [49]. However, the concern is that with AI it will be different because AI will be able to perform an increasingly wide range of complex tasks rather than being limited to unskilled or repetitive ones as previous technologies have been, thus driving the labor market toward a smaller and smaller subset of tasks for which human ability or presence is still required. In addressing this concern, it is first worth noting that economic measurements through the 2010s indicate that "[worries] about widespread disruption of the global labor market by AI have been premature" [12]. These economic findings are unsurprising given that much of the industrial sector is already automated, and the first wave of AI will serve to optimize automation that already exists. Nevertheless, Fathom5 believes it is likely that with continued innovation in AI over the coming decades, these worries may eventually be realized. This, we think, is an inevitable consequence of technological progress—we will eventually be able to make everything we need and more with very little labor.[15] Indeed, the prospect of a human population being free to pursue the goals and passions of its choosing with little concern for the essential necessities of life is a utopian one.

The question, then, is not how to prevent reduced need for labor in the long run but instead how to manage this transition in a way that does minimal harm. In the near-term, Fathom5 believes that the first applications of AI to automation should be in those industries for which there is the most to be gained, and there is a strong argument to be made for the industrial sector as an early adopter of AI-powered automation. Free same-day delivery of deodorant using automated drones is nice, but it is not a necessity. Finding a way to provide clean water, food, and energy more efficiently to a population of ten billion by 2050 is. In the long term, as automation becomes increasingly pervasive throughout society and the need for human labor continues to decrease, we as a society will need to confront the question of how to organize and distribute resources in a world where work has become largely obsolete.

Centralized Failure Mode

Another common concern regarding AI in the industrial sector is that it will introduce an additional, centralized failure mode for the operation of critical industries. Here, the fear is that a malfunction or cyberattack affecting an algorithm controlling core functionality in plants across a particular industry could cause harm at a societal scale by crippling our ability to produce and distribute essential goods and services. Imagine the harm that could be done if many electric plants across the globe simultaneously suffered catastrophic failures. However, Fathom5 believes that the risk of these outcomes is not made more likely by the addition of AI into the industrial sector and in fact can be reduced both directly and indirectly through the application of AI.

In the case of algorithmic malfunction, we point out that algorithms are already responsible for controlling critical industrial operations and have been for many decades, yet we have not seen the sort of large-scale failure outlined above. This is because the control systems deployed in the industrial sector are highly heterogenous, even across plants performing the exact same functions. Whereas one can write a single program that will perform the same function on any computer running Windows, no single algorithm can be deployed across plants performing a particular industrial function without a significant degree of customization for each plant. Therefore,

[15] It is interesting to note that economists as diverse as John Maynard Keynes and Karl Marx recognized nearly a century ago that continued gains in productivity would eventually lead to this so-called "age of leisure and abundance." John Maynard Keynes discusses this idea in his *Economic Possibilities for our Grandchildren*, while Karl Marx discusses these ideas across several works, in particular *Grundrisse*.

industrial accidents resulting from algorithmic malfunction are usually isolated events rather than industry-wide occurrences. This is true whether those algorithms are based on AI or manually programmed instructions.

The threat of widespread harm caused by cyberattacks is much more serious. Even for different algorithms designed by different firms, a malicious change to a particular input across plants could have disastrous consequences. Current industrial control systems are based on a design paradigm that was conceived long before the advent of the internet. In fact, many of these systems were designed before the term "cybersecurity" even existed.[16] The act of modernizing these systems to accommodate and take advantage of AI presents the opportunity to simultaneously improve cybersecurity as part of the first major overhaul of industrial control system architectures in over fifty years. This is central to Fathom5's "security first" approach to industrial automation. A modern industrial automation architecture must be born of a design philosophy that accounts for the potential of malicious actors taking advantage of the pervasive connectivity found within modern industrial systems. This is to say nothing of the potential of AI to monitor and deter cybersecurity threats, many of which are themselves increasingly enabled by AI. Therefore, the application of AI to the industrial sector, done right, has the potential to increase the cybersecurity of trillions of dollars in capitalized assets.

CONCLUSION

The harms outlined above should serve as a clarion call for the dangers of applying AI to problem domains without appropriate social and technical safeguards. Unfortunately, the reality is that developing these safeguards is likely to be a gradual process that unfolds over decades. In the meantime, the innovators that develop these technologies and bring them to the world will determine the trajectory humanity takes with respect to AI. Therefore, AI innovators have a moral imperative to carefully consider the implications of their actions—the evidence that these technologies can have negative consequences at scale is overwhelming, and there is no longer any excuse for thoughtlessly applying AI to problem domains that are causing widespread harm to individuals and our society.

As a company that seeks to benefit rather than harm humanity through the development and application of AI, Fathom5 has thought deeply about how AI can be applied responsibly in the current sociotechnical context. This thinking has led us to the industrial sector as an ideal sandbox where the application of AI can help tackle some of humanity's most critical challenges while posing minimal risk of widespread harm to individuals and societal institutions. We call on other innovators to do their own analysis to determine how their work on AI can best benefit humanity. AI is still an industry driven by people, and it will flourish wherever the people who work on it decide to apply their talents. If you are an innovator in the AI space reading this paper, you have just been personally implicated in this call to action, and your decision as to what to do next will determine the trajectory AI takes for years to come.

> **We call on other innovators to do their own analysis to determine how their work on AI can best benefit humanity. AI is still an industry driven by people, and it will flourish wherever the people who work on it decide to apply their talents.**

Authors

Brilliant Machines: A Pragmatic Approach to Responsible AI was written by Zachary Staples, founder and CEO, Fathom5 and Zachary Miller, Strategic Analyst, Fathom5.

If you would like to continue the conversation or explore partnership to create a new framework for responsible AI development, please reach out to ethics@fathom5.co

[16] The first recorded use of the term cybersecurity was in 1989 according to [50]

Military Disclosure

Fathom5 is proud to serve partners across the United States Department of Defense to demonstrate a technical and conceptual framework that ensures AI is developed and applied responsibly while achieving national security objectives.

Of course, there are serious ethical questions that arise when developing any technology deployed for a military purpose. Throughout our work, we apply the same ethical principles laid out in this paper – military applications of AI must begin with machine data for machine control in domains that are not fraught with the same ethical complications as military decisions about the use of force. For example, we are proud to develop a condition-based maintenance platform in support of maximizing the value of maintenance funds and extending the mean time between failure for industrial systems. Similarly, we support recognition of machine patterns that support improved identification of objects in a military area of operations.

Regarding matters of national security and the maintenance of global stability more broadly, it is critical that these questions be approached from a pragmatic standpoint. The reality is that malicious actors are willing to use all means at their disposal to advance authoritarian agendas, as the recent Russian invasion of Ukraine demonstrates. Given the paradigm shift in operational capabilities that will be brought by AI, the United States and our allies simply cannot afford to fall behind in AI innovation for military applications. Moreover, military history in the second half of the 20th century has shown that conflict tends to occur when there are asymmetries in military capabilities large enough to convince one side it can reliably achieve its objectives while incurring relatively few losses—this is the core principle behind the policy of mutually assured destruction that helped humanity avoid nuclear war. Indeed, every major military on earth is currently making significant investments to develop AI for military purposes. As such, global peace is most likely to be maintained if the United States military is viewed by its adversaries as an opponent with robust AI capabilities.

In an ideal world, AI wouldn't be developed for military applications—but this is not the world we live in. Whereas the decision to apply AI in the context of business is driven by opportunity, the decision to apply AI in the context of national defense is driven by necessity. Still, there are "sandbox" sectors of the defense sector that allow us to apply AI in a pattern recommended in this paper. Yet we also acknowledge that the ethical issues that arise in military applications of AI are not as clean as those in commercial applications. Ultimately, however, we believe that the development of AI for national defense is ethically justifiable as a real-world course of action aligned to our macro perspective that seeks to minimize the harm of AI in the context of warfare and reflects Fathom5's pragmatic approach to responsible AI.

REFERENCES

[1] E. Brynjolfsson and A. McAfee, The Second Machine Age. Norton & Company, 2014.

[2] T. Schleifer, "Google CEO Sundar Pichai says AI is more profound than electricity and fire," Vox, Jan. 19, 2018. Accessed: May 14, 2022. [Online]. Available: https://www.vox.com/2018/1/19/16911180/sundar-pichai-google-fire-electricity-ai

[3] "Existential Risk - Future of Life Institute." https://futureoflife.org/background/existential-risk/ (accessed Jun. 02, 2022).

[4] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Netw., vol. 2, no. 5, pp. 359–366, Jan. 1989, doi: 10.1016/0893-6080(89)90020-8.

[5] T. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 1877–1901. Accessed: May 24, 2022. [Online]. Available: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," Mar. 2015. Accessed: May 15, 2022. [Online]. Available: http://arxiv.org/abs/1412.6572

[4] Equal Credit Opportunity Act of 1974. 15 U.S.C. § 1691

[8] M. Hurley and J. Adebayo, "CREDIT SCORING IN THE ERA OF BIG DATA," Yale J. Law Technol., vol. 18, p. 69, 2016.

[9] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nat. Mach. Intell., vol. 1, no. 9, Art. no. 9, Sep. 2019, doi: 10.1038/s42256-019-0088-2.

[10] R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," Rochester, NY, Dec. 2017. Accessed: May 24, 2022. [Online]. Available: https://papers.ssrn.com/abstract=3086546

[11] C. F. Kerry, J. P. Meltzer, A. Renda, A. C. Engler, and R. Fanni, "Strengthening international cooperation on AI," Brookings Institute, Oct. 2021. [Online]. Available: https://www.brookings.edu/wp-content/uploads/2021/10/Strengthening-International-Cooperation-AI_Oct21.pdf

[12] Michael L. Littman, Ifeoma Ajunwa, Guy Berger, Craig Boutilier, Morgan Currie, Finale Doshi-Velez, Gillian Hadfield, Michael C. Horowitz, Charles Isbell, Hiroaki Kitano, Karen Levy, Terah Lyons, Melanie Mitchell, Julie Shah, Steven Sloman, Shannon Vallor, and Toby Walsh. "Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report." Stanford University, Stanford, CA, September 2021. Doc: http://ai100.stanford.edu/2021-report. Accessed: May 14, 2022.

[13] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault, "The AI Index 2021 Annual Report," AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2021.

[14] "Google says it's committed to ethical AI research. Its ethical AI team isn't so sure. - Vox." Accessed: May 14, 2022. [Online]. Available: https://www.vox.com/recode/22465301/google-ethical-ai-timnit-gebru-research-alex-hanna-jeff-dean-marian-croak

[15] "Facebook ad revenue 2009-2020," Statista. https://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/ (accessed May 14, 2022).

[16] M. Andreessen, "Marc Andreessen on Why Software Is Eating the World," Wall Street Journal. Accessed: May 14, 2022. [Online]. Available: https://www.wsj.com/articles/SB10001424053111903480904576512250915629460

[17] "Facebook: daily active users worldwide 2022," Statista. https://www.statista.com/statistics/346167/facebook-global-dau/ (accessed May 14, 2022).

[18] "Ledger of Harms." https://ledger.humanetech.com/ (accessed May 24, 2022).

[19] S. Madigan, D. Browne, N. Racine, C. Mori, and S. Tough, "Association Between Screen Time and Children's Performance on a Developmental Screening Test," JAMA Pediatr., vol. 173, no. 3, pp. 244–250, Mar. 2019, doi: 10.1001/jamapediatrics.2018.5056.

[20] J. S. Hutton, J. Dudley, T. Horowitz-Kraus, T. DeWitt, and S. K. Holland, "Associations Between Screen-Based Media Use and Brain White Matter Integrity in Preschool-Aged Children," JAMA Pediatr., vol. 174, no. 1, p. e193869, Jan. 2020, doi: 10.1001/jamapediatrics.2019.3869.

[21] P. G. Turner and C. E. Lefevre, "Instagram use is linked to increased symptoms of orthorexia nervosa," Eat. Weight Disord., vol. 22, no. 2, pp. 277–284, 2017, doi: 10.1007/s40519-017-0364-2.

[22] G. Wells, J. Horwitz, and D. Seetharaman, "Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show," Wall Street Journal, Sep. 14, 2021. Accessed: May 14, 2022. [Online]. Available: https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

[23] H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow, "The Welfare Effects of Social Media," Am. Econ. Rev., vol. 110, no. 3, pp. 629–676, Mar. 2020, doi: 10.1257/aer.20190658.

[24] E. Pickersgill, "Removed," Eric Pickersgill Studio. https://www.ericpickersgill.com/removed (accessed Jun. 02, 2022).

[25] Q. He, O. Turel, and A. Bechara, "Brain anatomy alterations associated with Social Networking Site (SNS) addiction," Sci. Rep., vol. 7, no. 1, Art. no. 1, Mar. 2017, doi: 10.1038/srep45064.

[26] Y. Sun and Y. Zhang, "A review of theories and models applied in studies of social media addiction and implications for future research," Addict. Behav., vol. 114, p. 106699, Mar. 2021, doi: 10.1016/j.addbeh.2020.106699.

[27] A. Hern, "'Never get high on your own supply' – why social media bosses don't use social media," The Guardian, Jan. 23, 2018. Accessed: May 14, 2022. [Online]. Available: https://www.theguardian.com/media/2018/jan/23/never-get-high-on-your-own-supply-why-social-media-bosses-dont-use-social-media

[28] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," Proc. Natl. Acad. Sci., vol. 110, no. 15, pp. 5802–5805, Apr. 2013, doi: 10.1073/pnas.1218772110.

[29] A. Dezfouli, R. Nock, and P. Dayan, "Adversarial vulnerabilities of human decision-making," Proc. Natl. Acad. Sci., vol. 117, no. 46, pp. 29221–29228, Nov. 2020, doi: 10.1073/pnas.2016921117.

[30] P. Lewis and P. Hilder, "Leaked: Cambridge Analytica's blueprint for Trump victory," The Guardian, Mar. 23, 2018. Accessed: May 14, 2022. [Online]. Available: https://www.theguardian.com/uk-news/2018/mar/23/leaked-cambridge-analyticas-blueprint-for-trump-victory

[31] G. Adomavicius, J. Bockstedt, S. Curley, and J. Zhang, "Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects," SSRN Electron. J., vol. 24, Dec. 2013, doi: 10.2139/ssrn.2285042.

[32] R. Epstein and R. E. Robertson, "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections," Proc. Natl. Acad. Sci., vol. 112, no. 33, pp. E4512–E4521, Aug. 2015, doi: 10.1073/pnas.1419828112.

[33] "IBISWorld - Industry Market Research, Reports, and Statistics." https://www.ibisworld.com/default.aspx (accessed May 14, 2022).

[34] "Total lobbying spending U.S. 2021 | Statista." https://www.statista.com/statistics/257337/total-lobbying-spending-in-the-us/ (accessed May 14, 2022).

[35] M. H. Ribeiro, R. Ottoni, R. West, V. A. F. Almeida, and W. Meira, "Auditing radicalization pathways on YouTube," in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, New York, NY, USA, Jan. 2020, pp. 131–141. doi: 10.1145/3351095.3372879.

[36] J. Horwitz and D. Seetharaman, "Facebook Executives Shut Down Efforts to Make the Site Less Divisive," Wall Street Journal. Accessed: May 14, 2022. [Online]. Available: https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499?mod=hp_lead_pos5

[37] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel, "Emotion shapes the diffusion of moralized content in social networks," Proc. Natl. Acad. Sci., vol. 114, no. 28, pp. 7313–7318, Jul. 2017, doi: 10.1073/pnas.1618923114.

[38] R. Fan, J. Zhao, Y. Chen, and K. Xu, "Anger Is More Influential than Joy: Sentiment Correlation in Weibo," PLOS ONE, vol. 9, no. 10, p. e110184, Oct. 2014, doi: 10.1371/journal.pone.0110184.

[39] A. J. Stewart, M. Mosleh, M. Diakonova, A. A. Arechar, D. G. Rand, and J. B. Plotkin, "Information gerrymandering and undemocratic decisions," Nature, vol. 573, no. 7772, Art. no. 7772, Sep. 2019, doi: 10.1038/s41586-019-1507-6.

[40] "Update on Twitter's review of the 2016 US election." https://blog.twitter.com/en_us/topics/company/2018/2016-election-update (accessed May 14, 2022).

[41] M. Isaac and D. Wakabayashi, "Russian Influence Reached 126 Million Through Facebook Alone," The New York Times, Oct. 30, 2017. Accessed: May 14, 2022. [Online]. Available: https://www.nytimes.com/2017/10/30/technology/facebook-google-russia.html

[42] M. M. Ribeiro and P. Ortellado, "Mapping Brazil's political polarization online," The Conversation. http://theconversation.com/mapping-brazils-political-polarization-online-96434 (accessed May 15, 2022).

[43] T. Marlow, S. Miller, and J. T. Roberts, "Bots and online climate discourses: Twitter discourse on President Trump's announcement of U.S. withdrawal from the Paris Agreement," Clim. Policy, vol. 21, no. 6, pp. 765–777, Jul. 2021, doi: 10.1080/14693062.2020.1870098.

[44] K.-C. Yang, C. Torres-Lugo, and F. Menczer, "Prevalence of Low-Credibility Information on Twitter During the COVID-19 Outbreak," p. 4.

[45] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[46] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," ProPublica. Accessed: May 14, 2022. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=XHXNgmPVImDaW15WrSaCeUz9xwCZGa7E

[47] S. Harris, "The Future of Artificial Intelligence," Waking Up.

[48] United Nations, "Population," United Nations. https://www.un.org/en/global-issues/population (accessed May 14, 2022).

[49] D. Acemoglu and P. Restrepo, "Automation and New Tasks: How Technology Displaces and Reinstates Labor," J. Econ. Perspect., vol. 33, no. 2, pp. 3–30, May 2019, doi: 10.1257/jep.33.2.3.

[50] A. Newtiz, "The Bizarre Evolution of the Word 'Cyber.'" Accessed: May 14, 2022. [Online]. Available: https://gizmodo.com/today-cyber-means-war-but-back-in-the-1990s-it-mean-1325671487

FATHOM5

BRILLIANT MACHINES