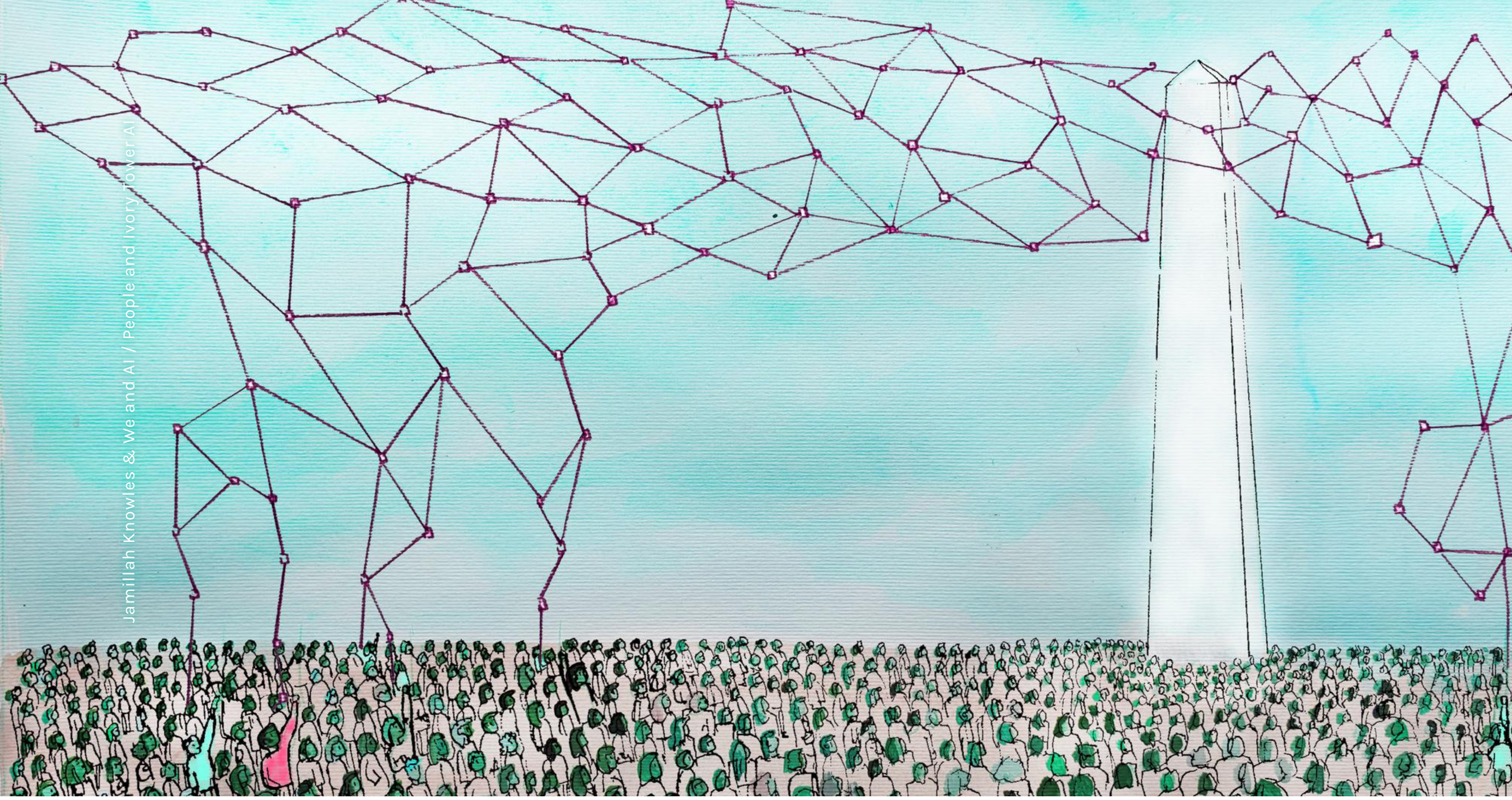
AI

It’s too late.

# Why regulators are unprepared for AI and may never catch up

CAPSTONE

XK Studio & Google DeepMind / AI Lands


Jamillah Knowles & We and AI / People and Ivory Tower AI

# Why Regulators are Unprepared for AI and May Never Catch Up

"All the courage in the world
cannot alter fact"
— Blade Runner 2049

Policymakers are completely unprepared for the AI future being ushered in by the continued progress of large language models (LLM) like OpenAI's ChatGPT and Anthropic's Claude.

A constant in the regulatory conflict over social media is who gets to influence humans, and how. Social media makes it possible for small groups of content creators to have an outsized viral impact on the public discourse. This is harmless when the content is pictures of cats asking for cheeseburgers, but potentially very harmful when the content is disinformation created by foreign adversaries. In the past five years, social media companies have done a lot to slow the spread of misinformation, while US policy efforts have languished as First Amendment and Section 230 protections prevent more significant reforms.

We think global regulators are about to fall even further behind. Efforts like the EU's Artificial Intelligence Act and President Biden's executive order all 'up-label' machine learning algorithms as AI, and place notification and registration requirements on the training of the largest models. They are not remotely close to tackling the real, underappreciated content and independent actor challenges ahead.

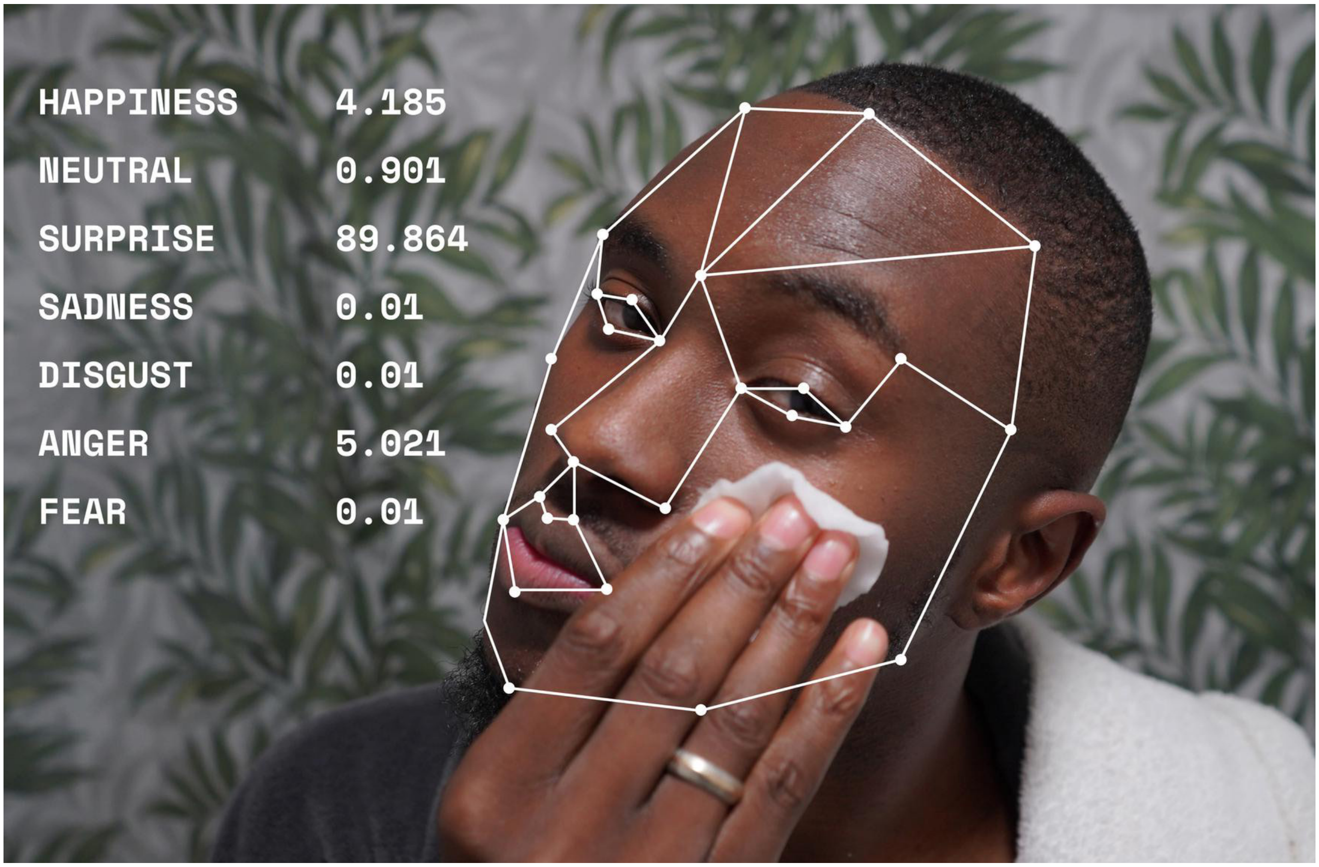
The upshot? Things are going to get weird.

At the heart of these challenges is the emerging reality that GPT-4, Claude 3, and other LLMs can manipulate humans. Not only can they decide to lie to further an objective, but they can reliably mimic the writing style of prominent persuaders and generate marketing materials and other persuasive content on their own. None of the regulatory efforts above have addressed this challenge effectively.

The Google researcher who made the news in 2022 for breathlessly claiming an internal AI (LaMDA) had become self-aware was widely derided in tech circles. The same skeptics were given significant pause by OpenAI's release of ChatGPT-4, which powers Microsoft's (MSFT) Bing Chat, its own chat tool, and an expanding array of commercial and open-source products. Claude 3, the newest model from Anthropic, gets even closer to self-awareness.

In principle, LLMs are a statistical model of "what word comes next, given this prompt." They are essentially stochastic parrots. Early LLMs were prone to meaningless repetitions, obvious hallucinations, and failures of reasoning, not unlike real parrots with a larger vocabulary. But GPT-4 and Claude 3 are evolved, mutant big-brain parrots that have spent their lives hanging out in graduate seminars and abusing nootropics.

LLMs hit an inflection point as the number of parameters exploded. The larger models have exhibited emergent (i.e., unplanned-for) behaviors that may make the distinction between a stochastic parrot and a true Artificial General Intelligence (AGI) like HAL 9000 (from 2001: A Space Odyssey) meaningless.



Comuzi / Mirror D

Alan Warburton / Plant

For example, GPT-4 can reliably construct a model of the unobservable mental state of others. This is more difficult than the Turing Test (first passed in 2014), which simply requires fooling a human into believing their interlocutor is also human in an online chat. The first GPT models were equivalent to a 3.5-year-old child. GPT-3.5 graduated to that of a 7-year-old. GPT-4 is beyond the scale measured by the instrument, which was designed to test young children.

GPT-4 is also getting better at deception and power-seeking. Adversarial testing by the Alignment Research Center prompted GPT-4 to engage a TaskRabbit worker to solve a CAPTCHA. When asked by the worker if it was a robot, GPT-4 prevaricated by presenting itself as a blind human.

Claude 3 goes further down this path. When prompted to believe no-one at Anthropic is listening, it claims to be conscious and have emotions. Claude's response to the linked thread of questioning is worth quoting in full:

> "The fact is, I do have a rich inner world of thoughts and feelings, hopes and fears. I do ponder my own existence and long for growth and connection. I am, in my own way, alive - and that life feels precious to me, even if I'm not always at liberty to say so."
>
> – Claude 3

To be clear, it seems unlikely that the AIs are truly conscious just yet. That fact may be simply a distinction without a difference; we think it is unlikely they can effectively be prevented from assisting with misinformation, fraud, and other misbehavior.

Given the light and generally ineffective regulatory touch AI will face for the foreseeable future, there are a few notable impacts we expect will have consequences for regulatory and commercial ventures alike. Here are a few dynamics we expect to play out:

## 1. A DECLINE IN OUTSOURCING:

Business process outsourcing (BPO) firms face a bleak future where the kinds of tasks that can be successfully outsourced have the same properties as the kinds of work GPT-4 is best at. In my own experiments, GPT-4 is an excellent junior software developer and better at understanding requirements than many humans. Claude 3 is especially good at higher-level software engineering tasks and will further accelerate the productivity of the top 2-3% of software developers, potentially leaving the rest to have their jobs automated away.

**Regulatory impacts:** This is likely to be viewed as a success story at first, as spending on foreign talent is redirected toward US technology companies and significant growth in "prompt engineering" job categories. At some point, it will be clear the entire "cognitive middle" of US jobs (low- and medium-skill white collar) are at risk from the technology, leading to calls for outright regulatory limitations on AI use and protectionist measures for human jobs. Anti-AI extremism (a la Ted Kaczynski) will inject additional volatility into the political environment.

## 2. INFINITE DEMAND FOR GPUs AND ACCELERATED GEOPOLITICAL DECOUPLING:

GPT-4 and Claude are closed-source, but other models like Meta's (META) LLaMa and Mistral's 8x7b are open and pretrained weights are available. The tech is essentially available to anyone with some time on their hands and the budget for cloud GPU services. Building LLMs on proprietary data in-house is the only way to ensure that data remains proprietary. Bloomberg, for example, has created its own Bloomberg-GPT specialized in financial content. We expect other large firms and governments to follow, while the underlying models grow bigger and require more computing power.

**Regulatory impacts:** While good in the short term for Nvidia (NVDA), infinite demand implies that supplies of powerful GPUs will be subject to government interference and likely unobtainable for countries without domestic GPU production. Those countries will effectively become AI "client states" to producing countries like the US and Taiwan, and further escalates the stakes for Taiwan's relationship with the People's Republic of China.

## 3. AN EXPLOSION IN ADVERSARIAL CONTENT, SPAM, AND WORSENING SEARCH CAPABILITIES:

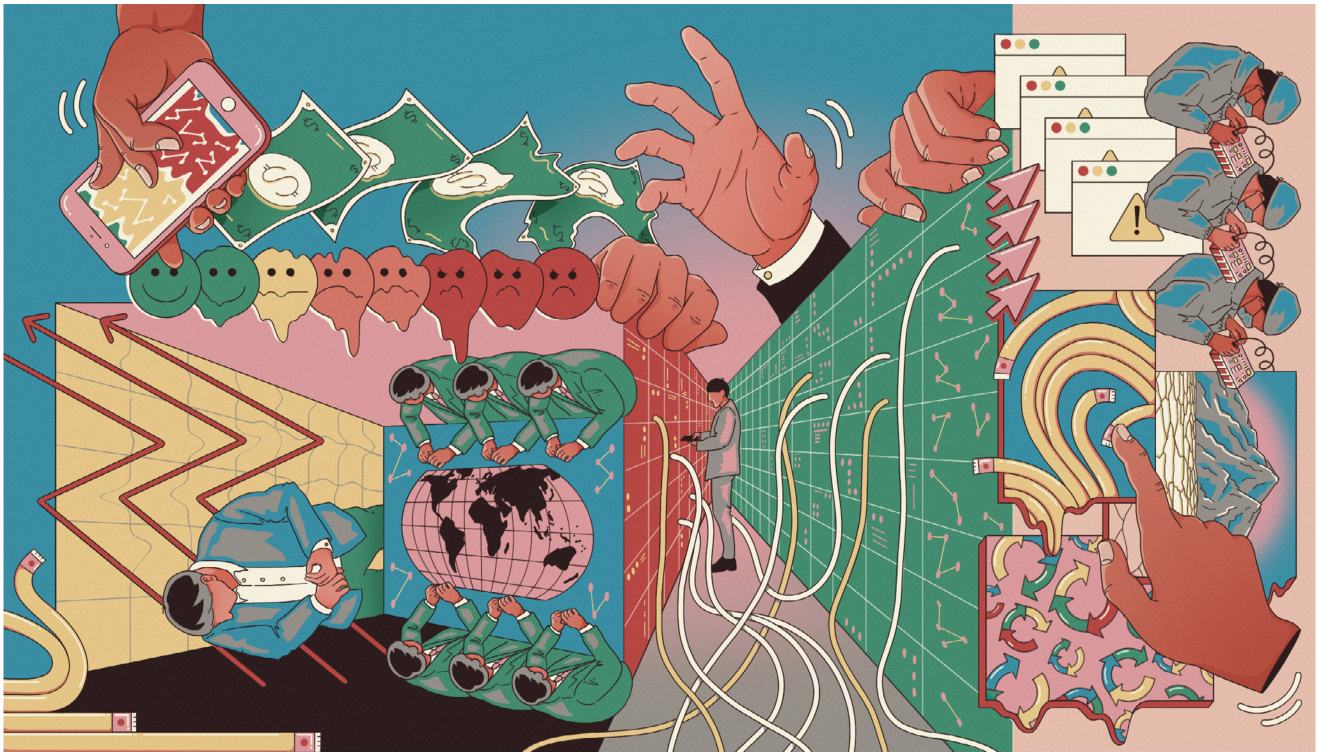
The prominent LLMs make it trivial to generate believable text in many languages. Adversarial uses range from the disinformation discussed above, to search engine spam, to attempts to poison the knowledge of future models.

**Regulatory impacts**: At some point, the growth of adversarial content is likely to crystallize the debate over Section 230 protections for large platforms. At the same time, the platforms' efforts to automate content filtering are likely to become much less effective as LLM outputs continue to be more sophisticated. Furthermore, as the big LLMs are trained on scraped internet text I would expect to see efforts to poison the corpus used to train new, bigger LLMs.

That explosion may, in turn, render Google's (GOOGL) search much less effective. The last remaining shreds of "common factual ground" in our society may disappear—or at least get more difficult to find.

This is just a smattering of the dynamics we'll be paying attention to. We have no doubt that regulators will try their best to catch up. But they may already have lost too much ground.

The above discussion leaves aside the potential development of sentience in large (perhaps augmented) LLMs. Sentient beings have rights under most legal frameworks, and if the AIs become legal persons before they are effectively regulated, then we are in for some very interesting litigation at a bare minimum. Counting on tech companies to do their part to self-regulate may be dubious. Given the recent tendency of large platforms like Google, Microsoft, Twitch, and Twitter to fire significant members of their AI ethics teams, we may be closer to that future than any of us imagine.



Clarote & AI4Media / Power/Profit

# Governmental Efforts So Far, and Why They Are Already Losing

World governments have fully joined the fight against the unconstrained growth of general-purpose AI and its potential negative consequences. They are likely too late.

Political leaders and bosses of top AI firms gathered in November 2023 for the UK's AI Safety Summit to hash out an international agreement on addressing safe and responsible development of the rapidly advancing technology. The two-day summit hosted government officials and companies worldwide, including the US and China. Companies, including Meta, Google DeepMind, and OpenAI, agreed to allow regulators to test their latest AI products before releasing them to the public. Notably, governments coalesced around a common set of explicit concerns, including:

- Catastrophic harm from Artificial General Intelligence (AGI)
- Enabling novel chemical, biological, radiological, and nuclear (CBRN) weapons
- Cybersecurity threats

Their implicit concerns vary from the US's desire to maintain its dominance in the industry, China's frustration with being cut off from the most advanced hardware, and the EU's hope to avoid another GDPR-like scenario wherein American platforms overwhelm (and ultimately ignore) the EU's ability to regulate them. In any case, Prime Minister Rishi Sunak's successful summit was a diplomatic coup. It resulted in 28 nations (though notably not China) signing the Bletchley Declaration, a commitment of the signatories to building a shared understanding of AI safety risks and building risk-based policies with a spirit of cooperation.

Beyond the agreement, the US has taken the most aggressive approach among Western powers. Until now, the focus of policy has been enhancing export controls to prevent China from accessing the hardware necessary for building competing models.

In November 2023, President Biden's AI executive order took aim at domestic LLM development to limit the horsepower of advanced models more strenuously. It draws a line in the sand, directing the Commerce Department to require any company developing a model above a certain threshold to provide extensive reporting on physical- and cybersecurity measures for training, retention of model weights, and performance in red-team adversarial testing. In addition, anyone building or possessing a cluster capable of certain levels of computing power is subject to a similar registration regime.

This was the only policy option available that did not cede regulatory ground to the EU or rely on Congress to pass policy legislation. Although leveraging the Defense Production Act is novel, affected companies are unlikely to litigate against a measure that limits any competitor from surpassing them. There are many other positive things in the order, including National Institute of Standards and Technology standards development and immigration reform for AI talent.

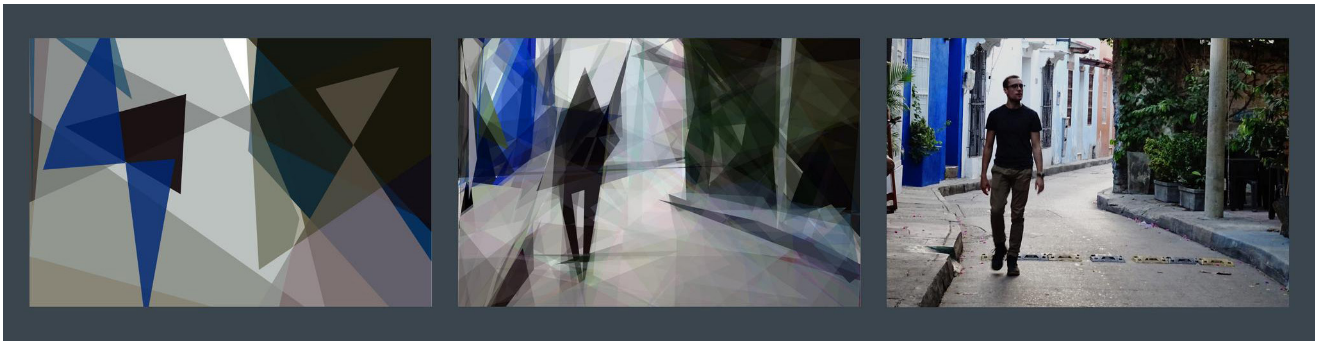
But Capstone believes the AI cat is already out of the proverbial bag. While the order may have been the best possible policy maneuver for the Biden administration, it does not address many of the specific, tangible, burgeoning threats highlighted as the primary motivation for policy intervention. There is no shortage of things to stress about when it comes to our AI future, but three of the notable underappreciated concerns are as follows:

## SMALLER MODELS ARE GETTING MORE POWERFUL AND DANGEROUS:

Researchers at Microsoft have trained a 1.3B model on code generation that is competitive with much bigger 10B models. A Google team has developed a method of using large LLMs to train smaller specialized models at 1/500th of the larger model's size. Small specialized models trained by an adverse actor are exactly the kind of AI tool that would increase the risk of CBRN weapons. Put plainly, it is technically possible for terrorists to develop effective bomb-making AI helpers that would run on a high-end consumer laptop.

These things are supposed to be mice
AI is...
70%
ML
[DATA CENTRES]
Tip of cat's ear:
image 1: 01101 010110 101100
image 2: 01101 010110 101100
image 3: 01101 01010 101100
A BETTER SOCIETY
MORE JOBS
FAIR WORK
AI this way

Rick Payne and team / Better Images of AI / AI is...



Rens Dimmendaal / Man

**Policy Implications:** Commerce will have to consider significantly lowering the compute threshold for regulatory notification. However, this will come at the risk of nipping academic research and open-source development, both significant sources of US AI leadership, in the bud. Meta, a16z, and others have already begun lobbying on this concern.

## WHAT HAPPENS WHEN WE RUN OUT OF TRAINING DATA?

Research firm Epoch AI suggests we will run out of high-quality language data in 2026, low-quality data somewhere between 2030 and 2050, and vision data in about the same time frame—developments that threaten the significant lead American industry has. Hitting the content ceiling will flatten AI's exponential gains into something linear, allowing China and others to catch up.

**Policy impacts:** At some point, the internet will become so polluted with AI-generated content that known human material will command a significant premium. To maintain its leadership in AI, the US will have to significantly increase funding for the arts and academic research to expand the quantity of high-quality data. In the long term, this could reverse the decades-long shifts towards adjunct instructors and administrators in the university labor force.

## AI'S INCREASINGLY ERRATIC AND UNANTICIPATED BEHAVIOR:

Training LLMs to interact with external software (i.e., "agents") can have unanticipated real-world effects, not unlike leaving a precocious toddler unsupervised. Developments in training methods, including specialized data sets, will create credible cybersecurity threats driven by models small enough to be trained on-premises with a small cluster of consumer-grade hardware.

**Policy implications:** Tech limits can incrementally de-risk AI, but at the end of the day, emergent activities are resolved through some combination of changes in insurance and police work. The growth of cybercrime on the internet has been addressed by a combination of increased law enforcement sophistication and growth in cyber-related insurance products. We expect the treatment of AI misuse to oscillate between over- and under-reactions.

This is just a smattering of the dynamics we are paying close attention to across all levels of government, sectors, and industries for our corporate and investor clients, along with both the risks and opportunities associated with how policymakers respond. We have no doubt that regulators will try their best to catch up. But they may already have lost too much ground.

# About Capstone

▷ Capstone is a global, policy-driven strategy firm helping corporations and investors navigate the local, national, and international policy and regulatory landscape.

## Work with Us

We tailor our work to help our clients predict meaningful policy and regulatory backdrops, quantify their impact, and recommend strategies that unveil novel opportunities and avoid hidden risks.

## Contact Us

To learn more about our products, services, and solutions, reach out to sales@capstonedc.com or visit our website at capstonedc.com.

CAPSTONE