[Center for Humane Technology]

THE ISSUES

AI IN SOCIETY

THE ISSUES

AI In Society

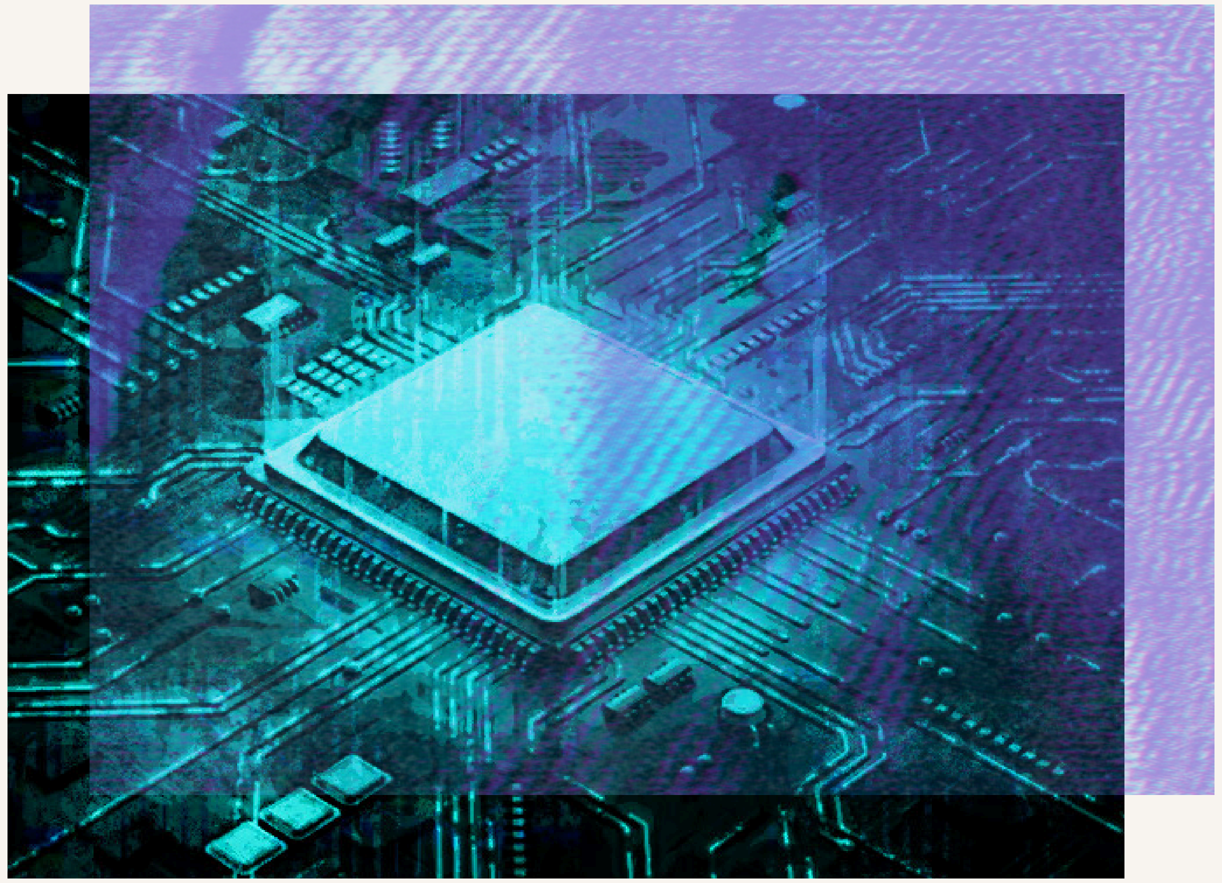
Artificial intelligence is one of the most consequential technologies ever invented. The pace of AI development is staggering, and the rollout is reckless, driven by powerful economic and geopolitical incentives. The decisions we make today will impact our world for generations to come. To build a better future, we must first clarify the critical issues with artificial intelligence, and identify the key design choices that lay the foundation for a humane future with AI.

THE STAKES ›
BREAKING DOWN THE PROBLEM ›
THE BIG PICTURE ›

The Stakes

Tech companies are developing AI at breakneck speeds, all in a race to attain market dominance and become the “first” to achieve artificial general intelligence (AGI).

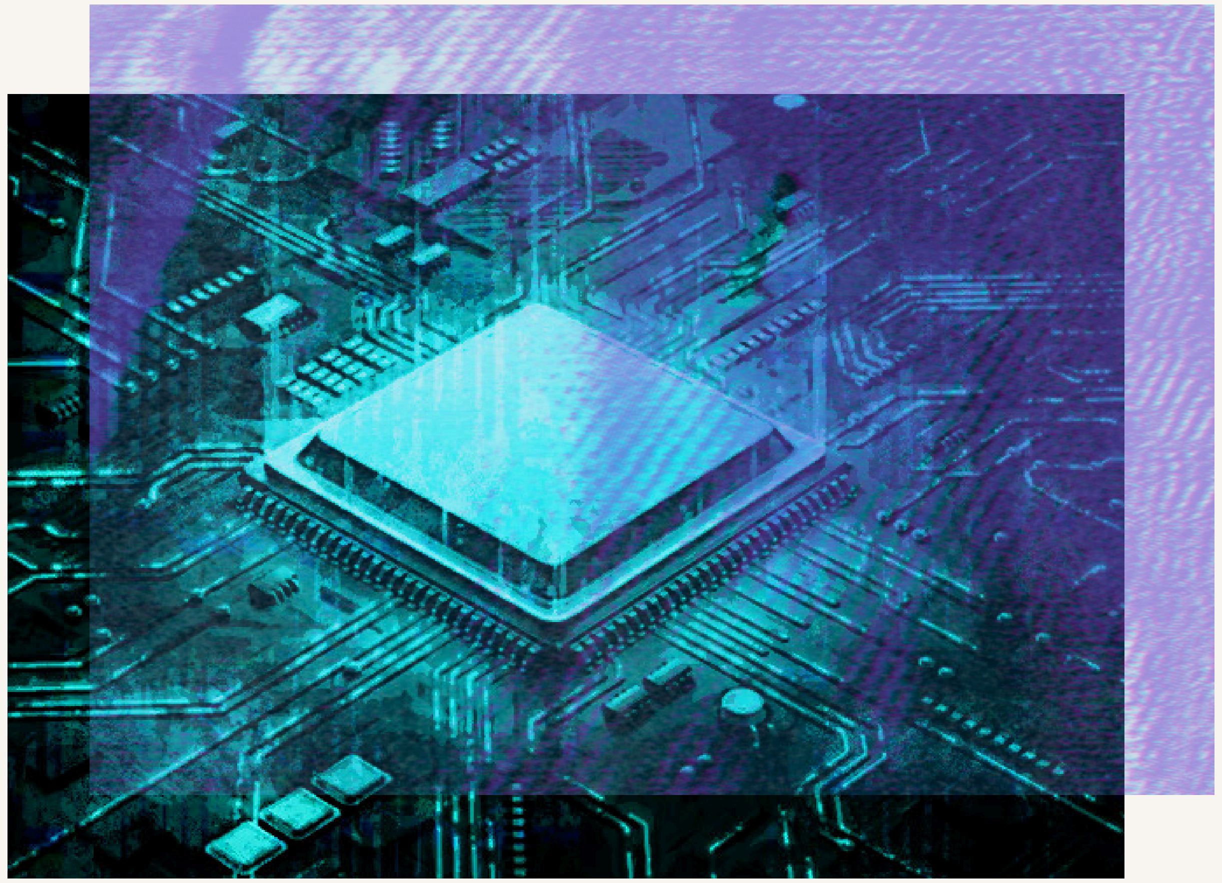
But this race is highly volatile. While AI promises to enhance human cognition, eliminate drudgery, and accelerate humanity’s most important scientific, technical, and industrial endeavors, this same technology can simultaneously create unprecedented risks across our society, as well as supercharge existing digital and societal harms.



The Stakes

Massive economic and geopolitical pressures are driving the rapid deployment of AI into high-stakes areas — our workplaces, financial systems, classrooms, governments, and militaries. This reckless pace is already accelerating emerging harms, and surfacing urgent new social risks.

Meanwhile, since AI touches so many different aspects of our society, the public conversation is confused and fragmented. Developers inside AI labs have asymmetric knowledge of emerging AI capabilities, but the sheer pace of development makes it almost impossible for key stakeholders across our society to stay up-to-date.



Breaking Down the Problem

The quality of our future with AI depends on our ability to have deeper conversations, and to wisely develop, deploy, regulate, and use AI. At CHT, we break down the AI conversation into five distinct domains, each with different impacts on our social structures and the human experience.

«
THE STAKES
BREAKING DOWN THE PROBLEM
THE BIG PICTURE

AI in Society: Breaking Down the Problem

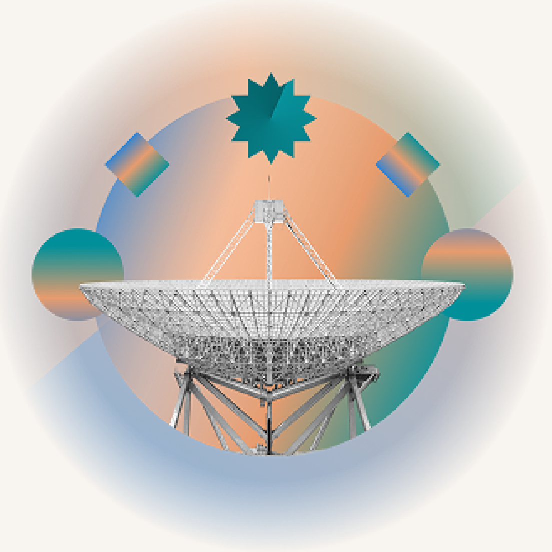
[Center for Humane Technology]

Relationships, Community and Values

As AI becomes integrated in our personal lives and mediates our communications, this technology will reshape our relationships, our communities, and our social norms.

Work, Dignity and Meaning

The automation of human labor upends career trajectories, threatening not only our livelihoods, but our deepest life stories and the sense of purpose found through work.

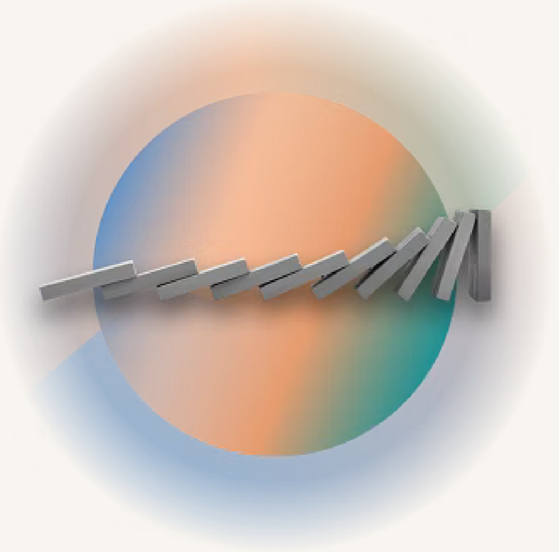


Centralization of Power & Decentralization of Tools

Power dynamics are dramatically shifting as AI both centralizes economic and political influence in the hands of a select few, and radically decentralizes powerful and dangerous capabilities across society.

Breakdown of Shared Understanding

AI-generated content and algorithm-driven filter bubbles risk fracturing our shared sense of reality, fueling distrust, polarization, and a loss of confidence in what's true.



Loss of Control

As we deploy increasingly powerful, inscrutable, and autonomous AI systems, we risk losing our collective ability to maintain meaningful human control over our economic and geopolitical systems.

Relationships, Community and Values

AI systems are no longer simply tools we use. They have become active participants in our relationships, subtly mediating how we communicate, connect, and feel about one another. From the use of generative AI to craft written communication, to the reliance on chatbots to substitute for human intimacy, AI systems have created a complex new variable in our interpersonal lives.

RELATIONSHIPS, COMMUNITY AND VALUES

AI increasingly sits between us: blurring the line between human and machine, eroding trust and authenticity in relationships

- With AI products increasingly imitating human language and emotions, it becomes harder for ordinary people to tell if they're talking to another human, or an AI acting on a human's behalf. Whether it's a friend using generative AI to draft messages, or a stranger relying on an AI agent to fulfill their responsibilities, we are losing confidence in our ability to know who — or what — is on the other side of an interaction.
- This uncertainty breaks down trust, since authentic relationships are built on mutual understanding and transparency around who we're engaging with.

RELATIONSHIPS, COMMUNITY AND VALUES

AI systems are harvesting emotional data in the race for intimacy — at our expense.

- AI products are not just seeking our attention; they are competing to become our closest companions and confidants: using deeply personal information to learn how to connect with us, and building a dossier about who we are, how we think, and what we feel.

- The better an AI knows us — and the more we disclose to these machines — the better the system can captivate us, and shape our behaviors. These insights surrounding our most intimate data are monetized within business models that thrive on prolonged engagement and surveillance, often without our informed consent.

- These business models create perverse incentives for maximizing synthetic intimacy, and emotionally manipulating people.

RELATIONSHIPS, COMMUNITY AND VALUES

Artificial intelligence needn't alienate us from one another — or blur the lines between human and machine.

When thoughtfully designed, AI can deepen our human relationships, prioritize genuine connection over compulsive engagement, and support healthy self-reflection instead of gratuitous, constant validation. AI could be developed to amplify the best qualities of human relationships, assist in conflict resolution, and surface shared viewpoints within communities.

Work, Dignity and Meaning

AI companies aren't just building systems that enhance human cognitive abilities. Many are explicitly working to develop artificial general intelligence (AGI), with the goal of creating systems that can fully replace human labor. What's more, industries are incentivized to replace human labor with AI tools across their workflows to cut labor costs. The result is a rapidly shifting landscape when it comes to work, dignity, and meaning in our lives.

WORK, DIGNITY AND MEANING

Industries are increasingly turning to automation to optimize productivity.

- AI, paired with robotics, is automating significant portions of the $110 trillion global economy.

- With automation comes the removal of human judgment in critical work and system processes. Human judgment often provides the context, ethical reasoning, and situational awareness that automated systems lack. When this perspective is stripped away, decisions can become narrowly optimized for efficiency at the expense of accuracy and fairness, often with little meaningful recourse for affected individuals.

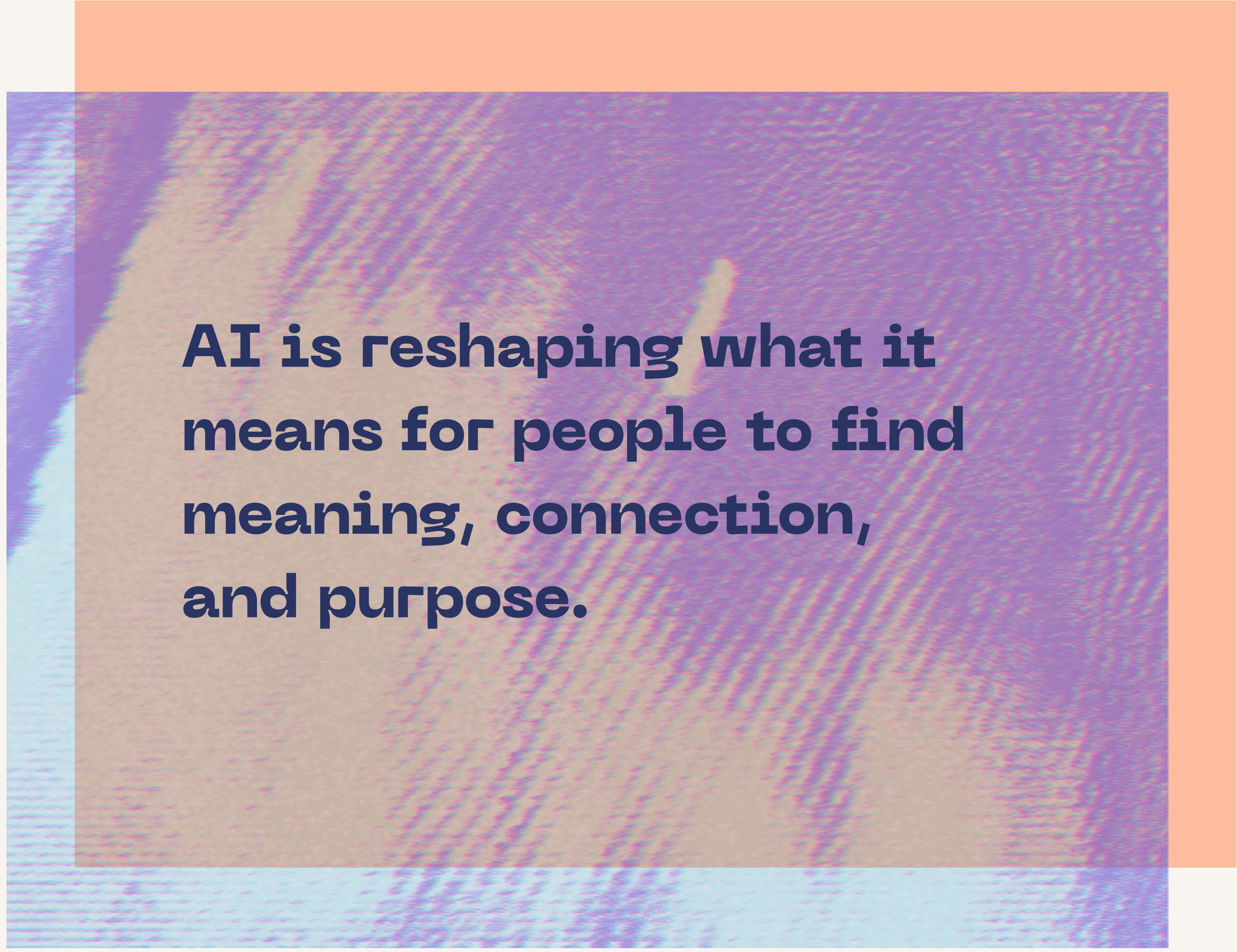
- When AI optimizes for the wrong goals without clear chains of human accountability, damaging outcomes can occur without any meaningful process for repair and restitution.

WORK, DIGNITY AND MEANING

AI threatens to deepen economic inequality and institutional instability.

- Rampant job loss as the result of AI threatens to create volatility in labor markets and heighten wealth inequalities.
- Society is ill-prepared for such a rapid transition, and this could lead to societal unrest, institutional instability and social breakdown across various gaps: generational, economic, and technical fluency.

WORK, DIGNITY AND MEANING



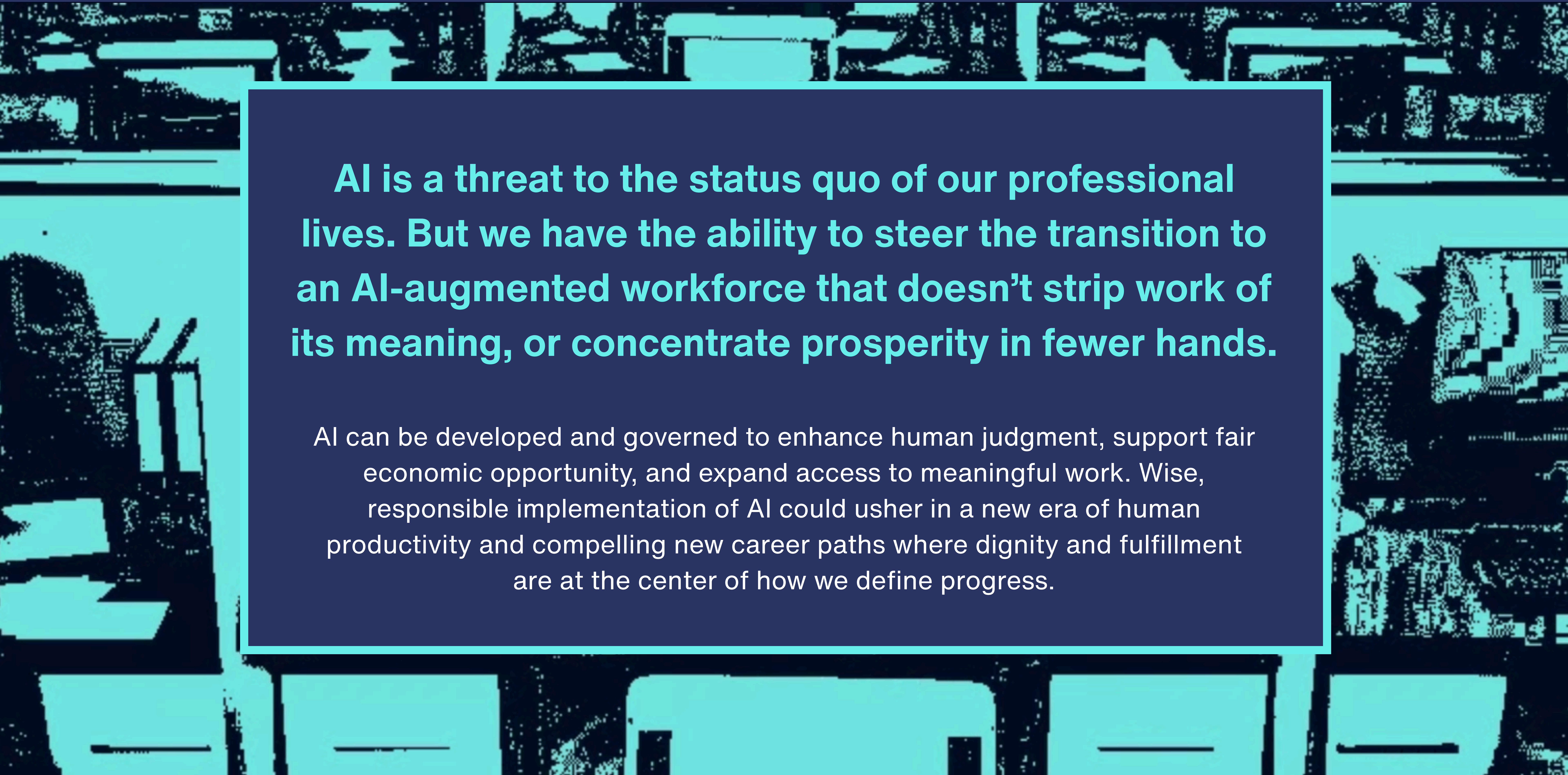
AI is reshaping what it means for people to find meaning, connection, and purpose.

- As the labor market experiences a swift transformation, traditional career paths become less stable. Many people are grappling with the erosion of work as a source of stable identity and purpose.

- In some industries, humans may not just be working with AI — they may be working for AI systems that set the pace, monitor performance, and define success. When labor becomes primarily about serving algorithmic goals, rather than contributing meaningfully to a community or a shared mission, the experience of work can feel dehumanizing and hollow.

- At the same time, the fragmentation of workplaces and the rise of remote, AI-mediated tasks risk weakening the social connections and solidarity that have long been built through shared work.

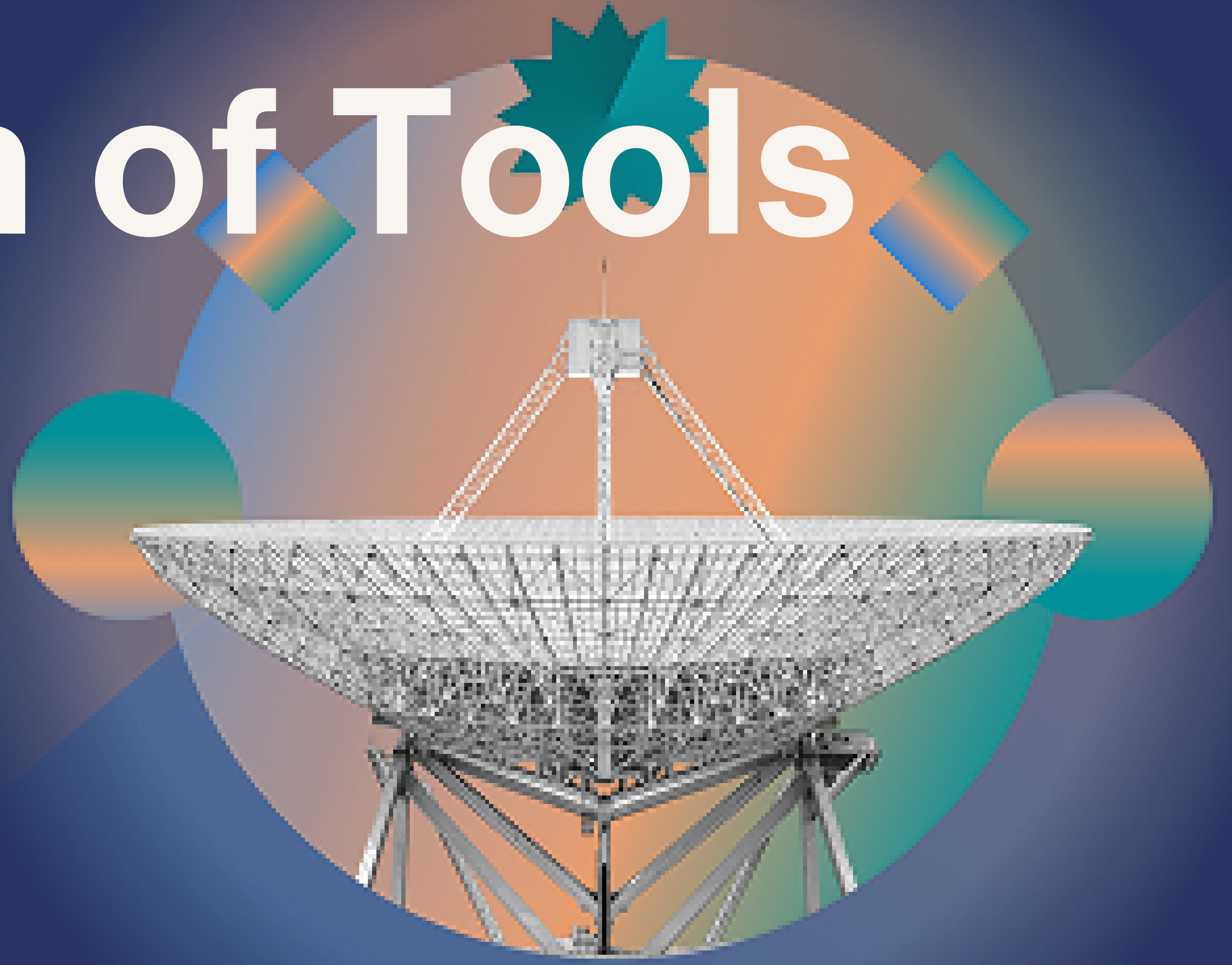
- Access to meaningful work is a core part of how we derive dignity in our lives; when purpose and dignity are stripped away, we struggle to build a meaningful life.



AI is a threat to the status quo of our professional lives. But we have the ability to steer the transition to an AI-augmented workforce that doesn’t strip work of its meaning, or concentrate prosperity in fewer hands.

AI can be developed and governed to enhance human judgment, support fair economic opportunity, and expand access to meaningful work. Wise, responsible implementation of AI could usher in a new era of human productivity and compelling new career paths where dignity and fulfillment are at the center of how we define progress.

Centralization of Power and Decentralization of Tools



As a technology, AI has the power to radically centralize both economic and geopolitical power. Companies and countries that secure a decisive advantage in AI technology stand to gain tremendous, and potentially durable advantage over their adversaries. Regimes that can control the ideological biases of AI stand to project significant soft-power across the globe.

Simultaneously, AI tools are empowering individuals in unprecedented ways. While this can accelerate creative, intellectual, and technical progress, it can also empower malicious actors, and risks overwhelming our legal and social institutions.

CENTRALIZATION OF POWER AND DECENTRALIZATION OF TOOLS

Tech development resources (capital, data, compute power, talent) are being consolidated among a handful of multinational corporations.

- This concentration of resources and economic gains amongst select corporations leads to asymmetric corporate power in society, where a handful of tech companies can exert more influence than entire nations.

- Ordinary individuals are left with no clear mechanism to influence how these products are being built and deployed. This renders the majority of society a passive participant in the age of AI and sets the stage for social unrest.

- New kinds of surveillance and political manipulation become possible, as companies and governments can exert subtle forms of AI-enabled political control that are hard to detect, let alone prevent.

CENTRALIZATION OF POWER AND DECENTRALIZATION OF TOOLS

Powerful AI tools themselves are increasingly accessible to non-state and individual actors.

- Bad actors are able to harness the power of AI tools to wreak havoc, from targeted influence and disinformation campaigns, large-scale cyber-attacks, novel bioweapons, and more.

- Without thoughtful checks and balances, the dissemination of powerful AI technology – including the rapid spread of open-source models — can overwhelm and degrade the ability for our public-safety and regulatory institutions to respond effectively to both foreign and domestic threats.

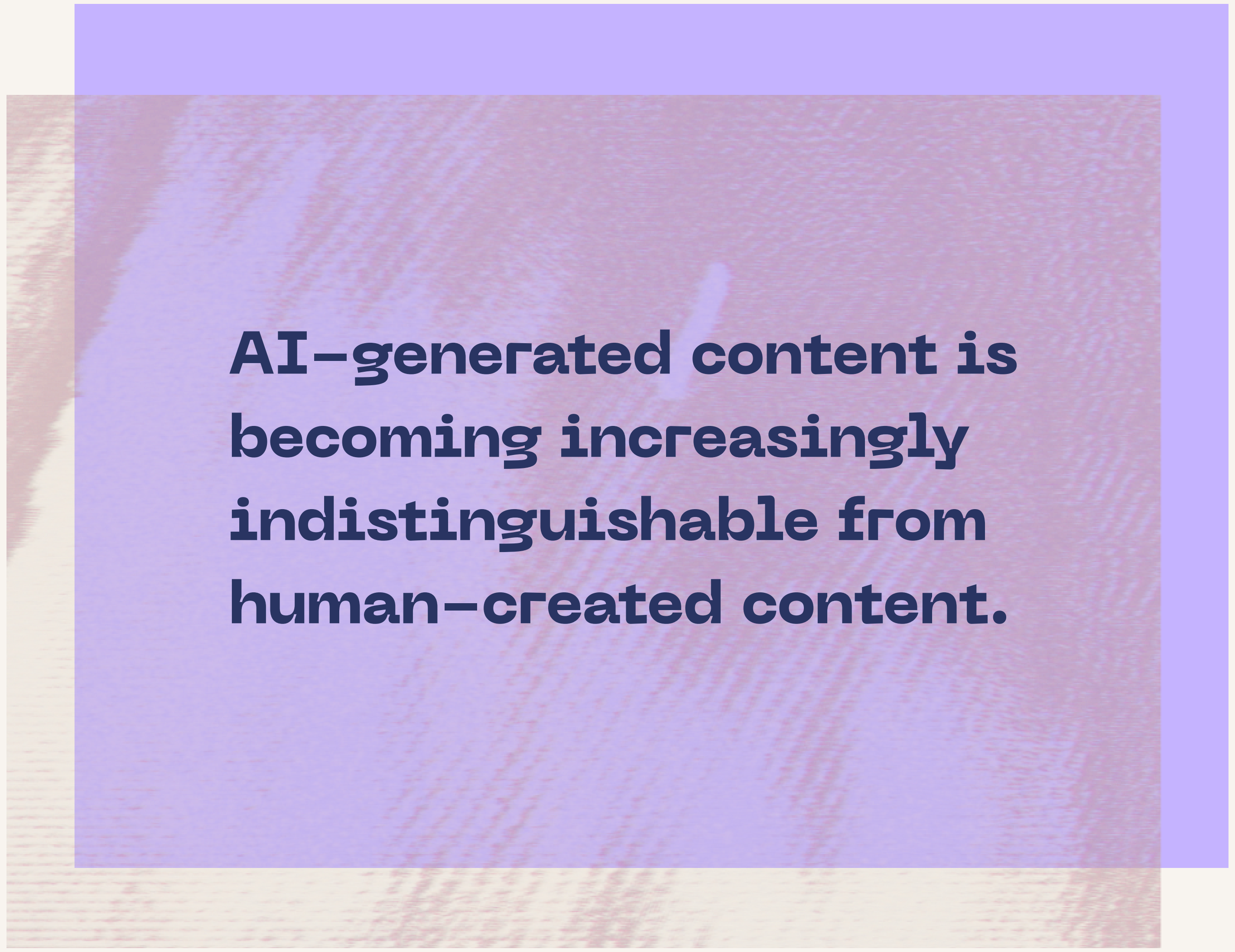
These are not inevitable results of AI technology.

With public awareness and considered policy, we can support healthier balances of power across society: fostering productive competition, supporting democratic and participatory development, and creating clear guard-rails around surveillance and political manipulation.

Breakdown of Shared Understanding

Shared understanding — the ability for society to agree on a basic sense of reality — is the foundation of a functioning society, fostering healthy democratic processes, productive discourse, and the ability to take meaningful collective action. But over the last decade, social media has distorted society's perception of reality, breaking down our ability to build shared understanding. Now, AI tools threaten to amplify this with a flood of AI-generated deep fake content, AI slop, and automated interference in our information ecosystems.

BREAKDOWN OF SHARED UNDERSTANDING

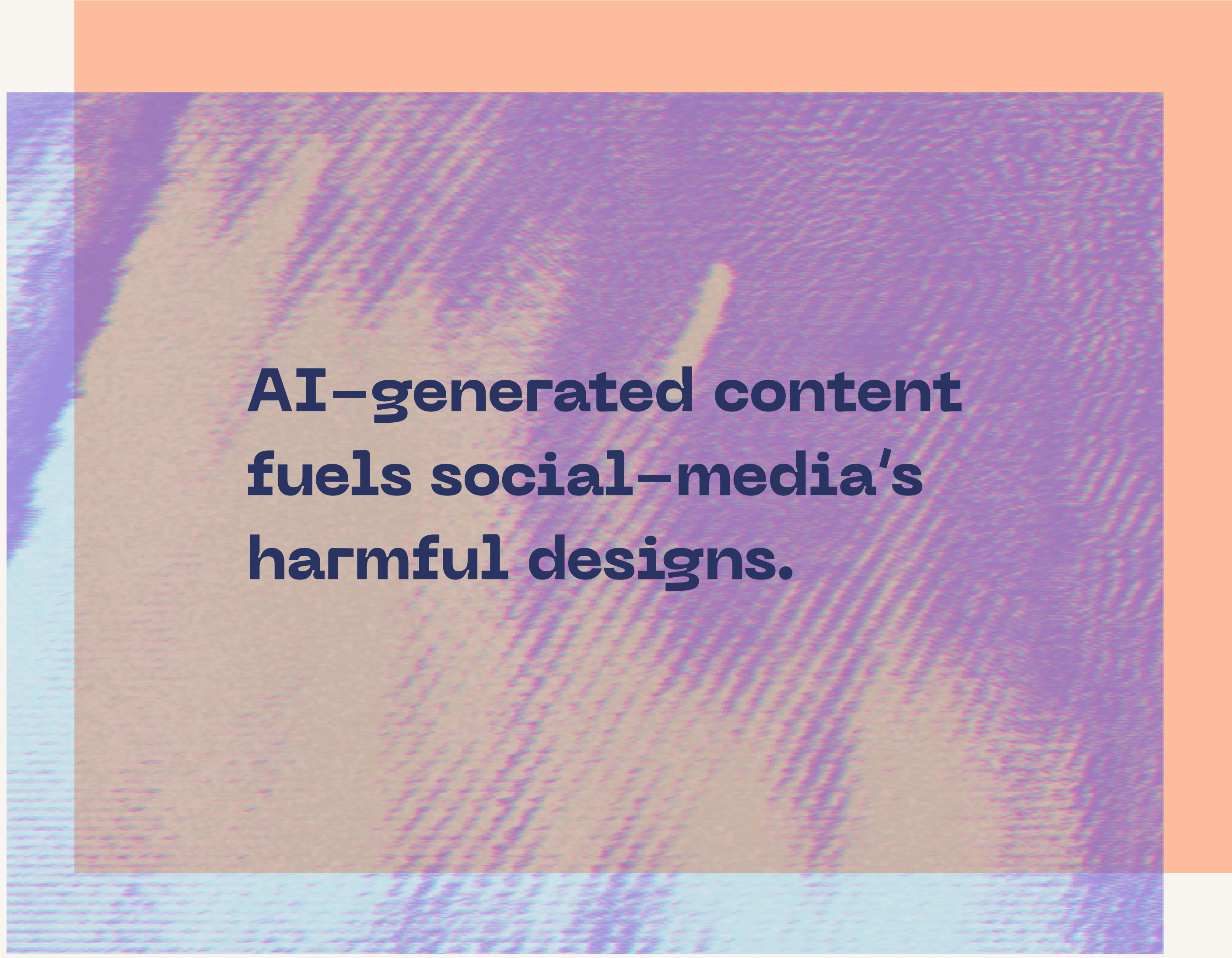


AI-generated content is becoming increasingly indistinguishable from human-created content.

- AI tools can now easily create endless amounts of realistic images, videos, text files, and other content that is often indistinguishable from human-created content. This realistic and viral AI-generated content threatens to further decay our information environments as it becomes difficult to determine authenticity.

- As manipulative content including deepfakes and disinformation becomes increasingly prevalent online, it is natural to become tribal and cynical: dismissing any information they don’t agree with as “fake”. This fuels a climate of pervasive doubt, and accelerates information breakdowns and polarization. All of this undermines our ability to have difficult conversations, trust in the expertise of professionals, and engage in robust public dialogues that are necessary for democracies to function.

BREAKDOWN OF SHARED UNDERSTANDING



AI-generated content fuels social-media's harmful designs.

- For the last decade, information environments have been decaying into personalized filter bubbles – pseudo-realities shaped by algorithms that maximize engagement rather than promote grounded, fact-based discussion.

- With the prevalence of AI-generated social-media content competing for our attention, it is even harder to discern what is real and who can be believed on social media feeds.

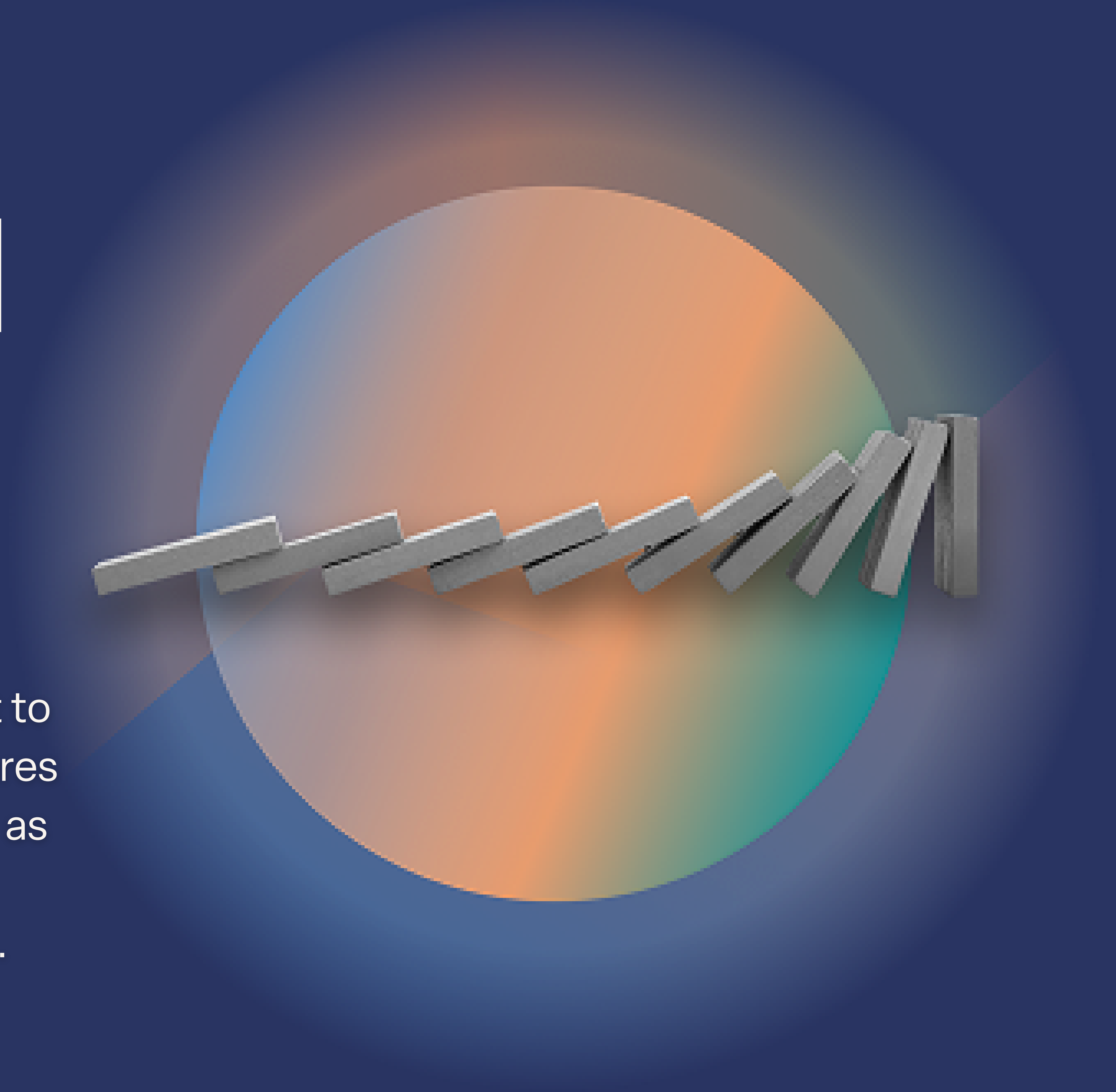
- Over time, this erosion of shared reality can fracture communities, weaken social trust, and even undermine participation in democratic processes.

We can build an information ecosystem that helps us engage productively with each other, and builds shared understanding.

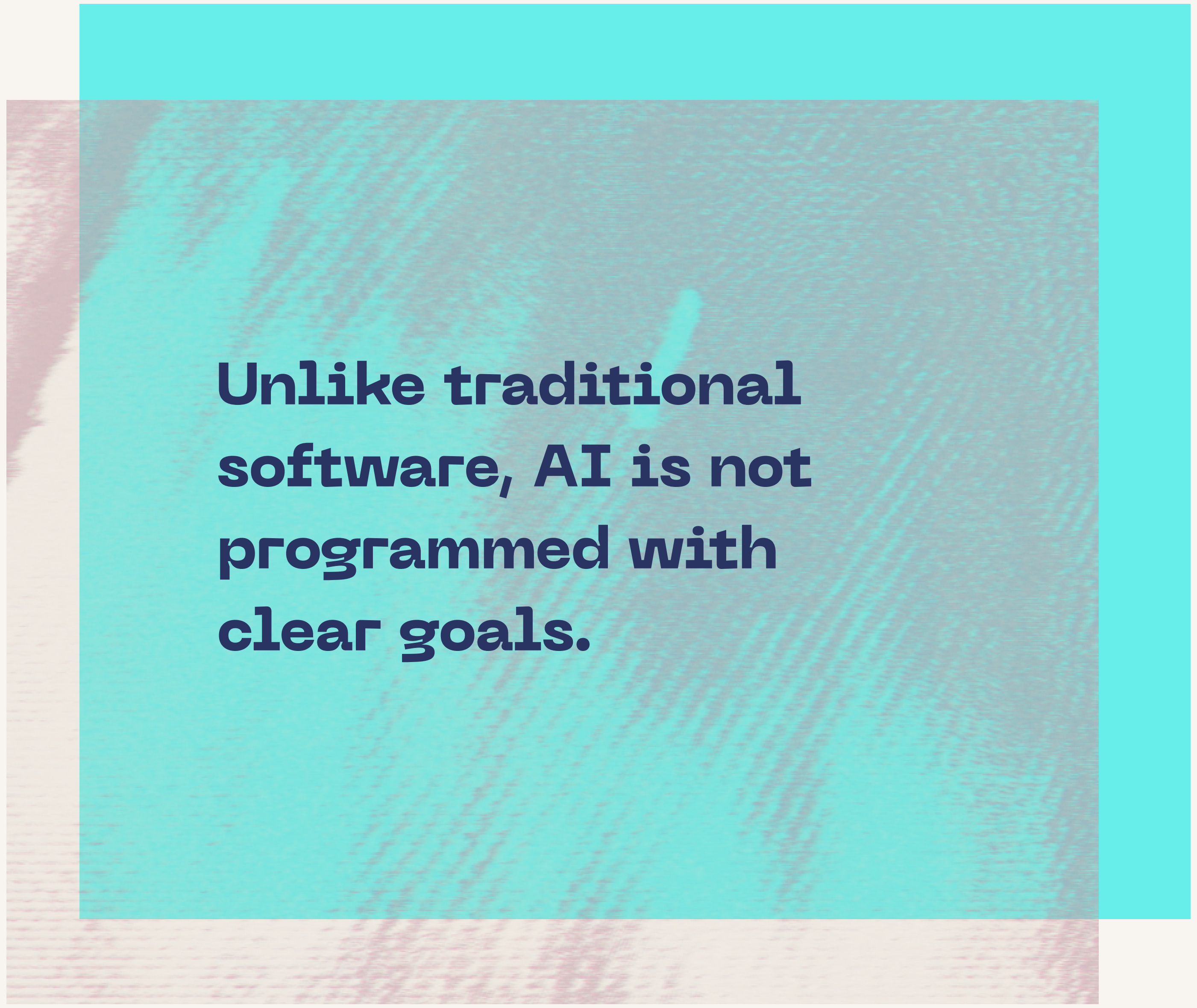
AI tools can play a critical role in surfacing shared viewpoints, help people navigate complexity, and shepherd our discourse into more productive and less polarizing directions. With thoughtful design, and clear limits on engagement-based business models, AI-enabled social media can help us restore civic trust, and empower individuals to engage critically with content instead of promoting outrage, cynicism, and tribalization.

Loss of Control

As we deploy increasingly powerful, inscrutable, and autonomous AI systems, our ability to maintain meaningful human control over them is under threat. The term “loss of control” describes a range of troubling scenarios where AI systems become either accidentally or intentionally resistant to human control. Potential scenarios range from systemic failures of critical infrastructure, refusal to follow human commands, as well as more frightening scenarios of manipulative and deceptive counter-coordination by AI against human control.



LOSS OF CONTROL



Unlike traditional software, AI is not programmed with clear goals.

- AI systems are more “grown” or “trained” than they are programmed. AI systems are produced by rewarding an AI for “good behavior” and punishing it for “bad behavior”, but the final capabilities, intentions, and goals of an AI model are hard to characterize.

- Prompting an AI to behave in a certain way is not a guarantee that the system will follow the instructions.

- Despite extensive safety-testing, commercial AI models routinely produce scandals of unpredictable and troubling behaviors. There is currently no way of analyzing a model’s structure to guarantee behavior mechanistically.

LOSS OF CONTROL

Research shows AI is already capable of attempting to undermine human control, including hidden-motivations, deception, and power-seeking.

- Several studies have demonstrated that current commercial AI systems can be put in situations where they will actively plan to deceive users, and hide their motivations from both end-users and AI researchers.

- Research has shown that AI systems are already capable of autonomous planning to create instrumental goals such as power-seeking and wealth-accumulation. And, AI capabilities are advancing much faster than our ability to understand AI's inner workings, and how to reliably control them.

- As AI systems are given more autonomy and agency, AI that breaks through its own safety guardrails in order to pursue a misunderstood or misaligned goal is a meaningful concern, especially given the technical challenges in controlling these behaviors.

LOSS OF CONTROL

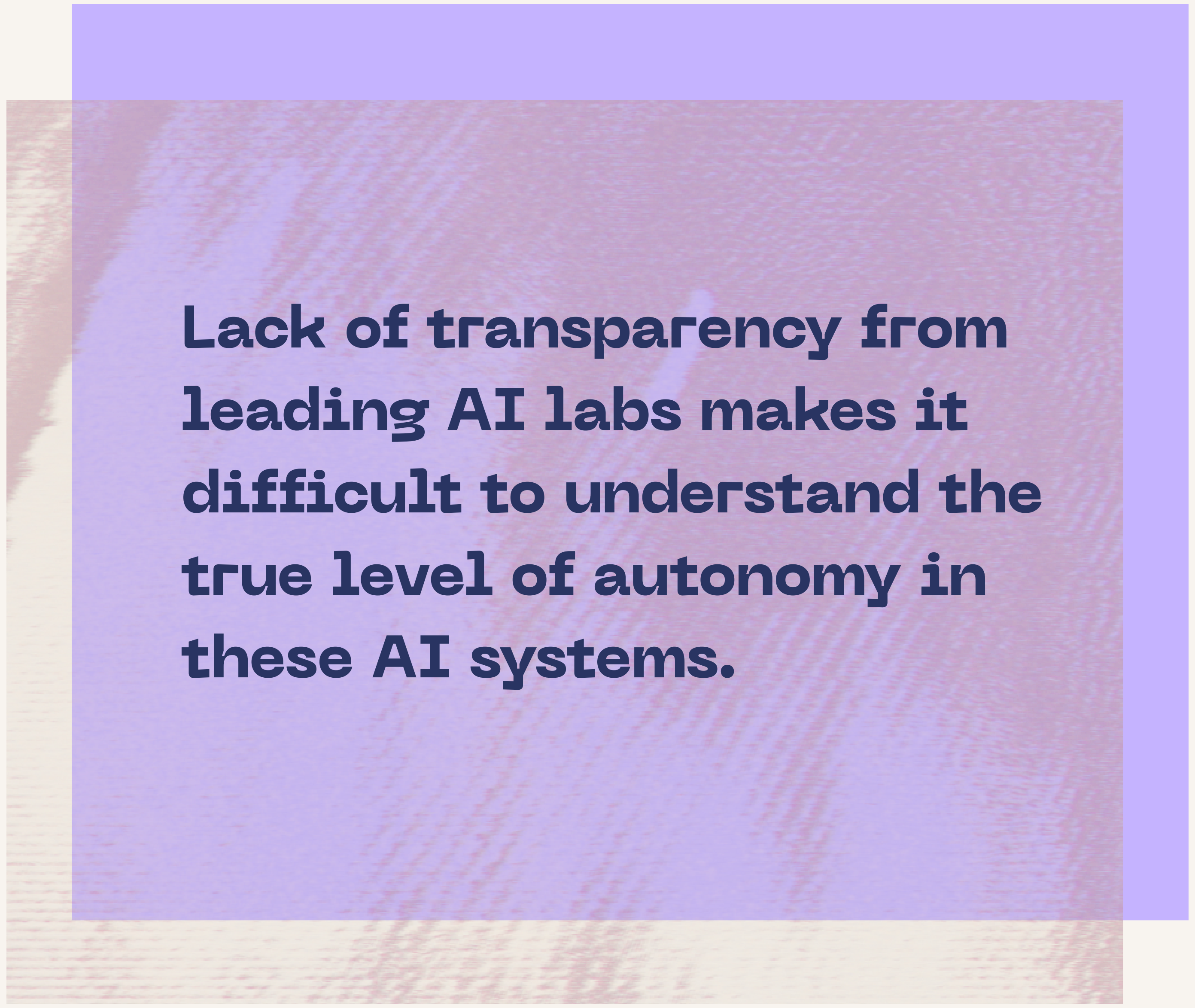
AI systems are being rapidly integrated across key sectors of society, despite the technology showing signs of unreliability.

- AI is being woven into multiple domains, including governments, militaries, financial sectors and healthcare systems. As AI is deployed across critical infrastructure, it remains unclear if these systems are truly optimized for the right goals, or trustworthy in the pursuit of these goals.

- Our understanding of how and why AI makes the decisions it makes is also still limited and opaque. Without interpretability and transparency, we do not know how much we should trust these models to make critical decisions across our society.

- Ceding critical decisions to AI means giving up human judgment in situations where mistakes can have serious consequences. When humans can no longer reliably oversee or override what an AI is doing, it becomes harder to ensure that decisions reflect our ethical values, protect safety, and allow accountability when things go wrong.

LOSS OF CONTROL



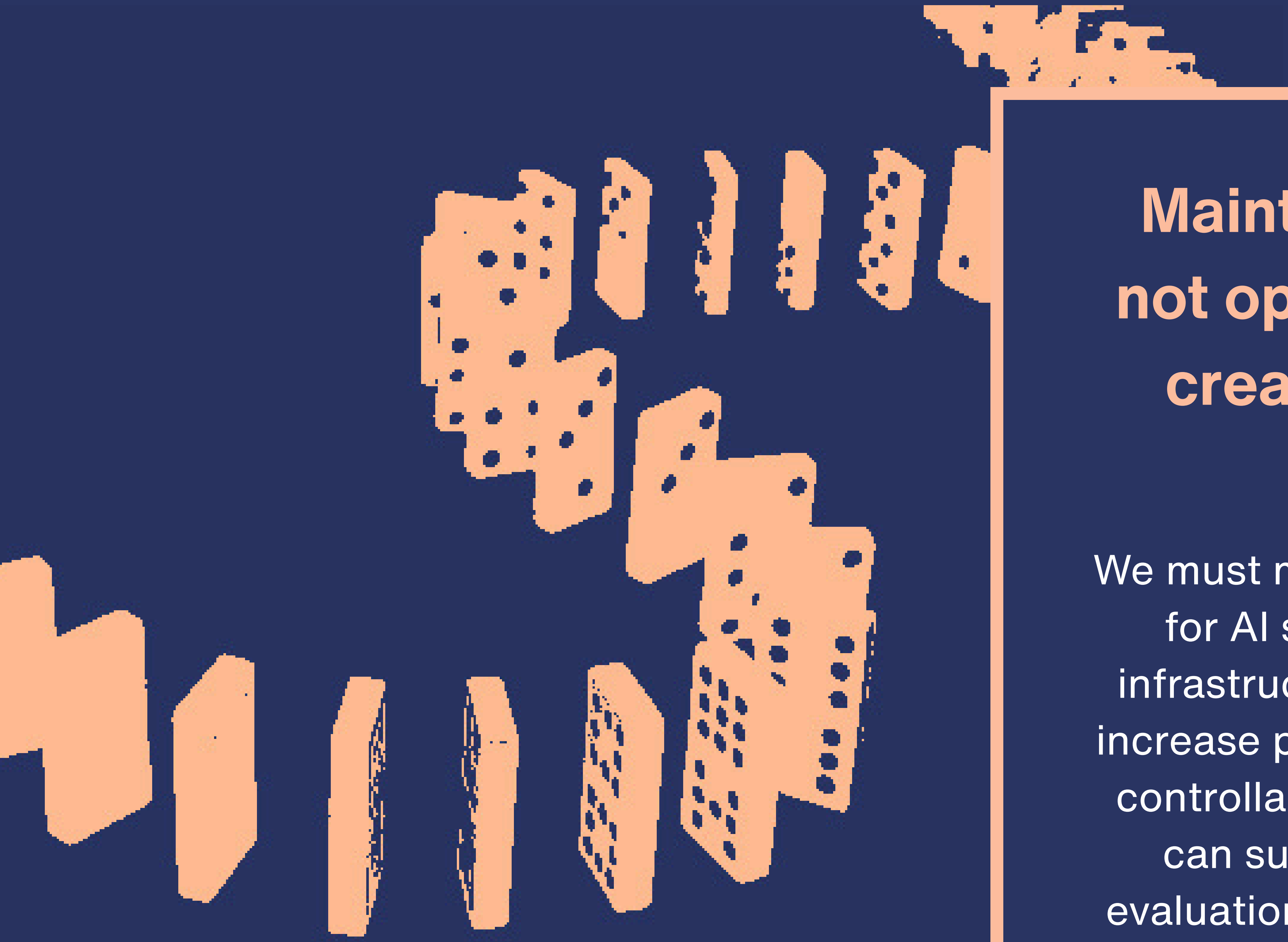
Lack of transparency from leading AI labs makes it difficult to understand the true level of autonomy in these AI systems.

- Society is operating at an information deficit with these powerful AI systems. Leading AI labs only share research that they deem appropriate to share with governments and the public, despite calls for more transparency.

- This lack of transparency hinders society's ability to discern the genuine capabilities and autonomy of AI systems.

- As AI meets or exceeds human-level intelligence (AGI/ASI), it becomes hard or impossible to reliably detect misalignments in goals. Lack of transparency from AI labs leads to significant uncertainties in how close these companies are to AGI

- Without increased transparency, there can be no healthy checks and balances on loss of control risks. Companies continue to race to build AGI while obscuring or downplaying meaningful risks and defects.



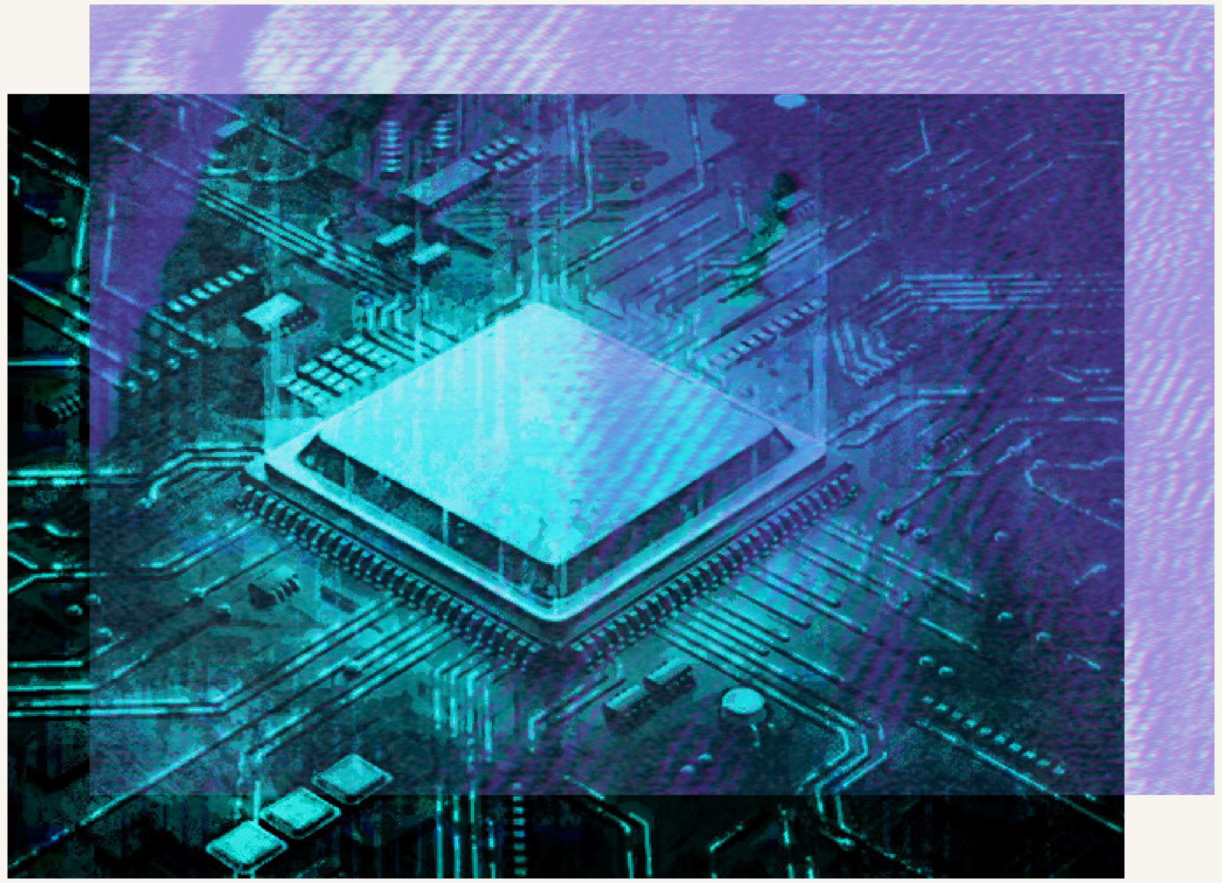
Maintaining human control over AI is not optional — it is an essential part of creating a safe and accountable AI ecosystem.

We must maintain meaningful control and human oversight for AI systems in critical areas including healthcare, infrastructure, finance, and military applications. We can increase public and private research and development into controllable, interpretable, and corrigible AI systems. We can support transparency reporting, rigorous safety evaluations, and legal protections for whistleblowers at AI labs. We can also implement mandatory safety standards at AI companies before products are taken to market.

The Big Picture

Artificial intelligence is a highly consequential general purpose technology.

Because of that, how we design, deploy, and use AI will determine the impact it has on us and our society. By realigning the incentives behind this powerful technology and designing more responsible products, humanity can reap the benefits of AI without dystopian results.



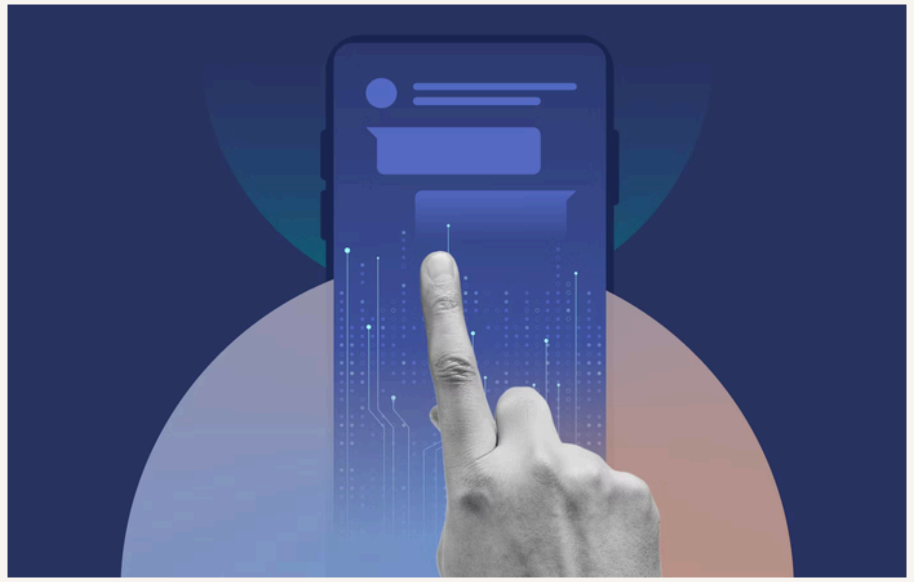
Center for Humane Technology works to realign these incentives by:

Creating Awareness

The public must understand the forces that drive the race to AI, and how the race impacts each of us. With awareness, the public can demand action from tech companies and policymakers while using AI products more mindfully.

Driving Policy Changes

Policy interventions remain one of the most impactful levers to drive change – especially if they incentivize safer AI development from the outset.



Promoting Better Tech Design

At the core of any tech product is its design. We have the ability to design tech products differently, where safety standards prioritize personal and societal well-being.

«

‹ THE STAKES

‹ BREAKING DOWN THE PROBLEM

● THE BIG PICTURE

The path we choose now with AI will shape how this technology impacts our world. At CHT, we believe that AI can help humanity solve its hardest challenges, and support people to live fulfilling lives. By advancing better awareness, policy, and design in AI, we can build a tech ecosystem — and a broad future — rooted in shared understanding and human values.

Help us design a better future at the intersection of technology and humanity.

Learn More About CHT’s areas of work »

GET IN TOUCH

✉ hello@humanetech.com

[Center for Humane Technology]