THE DIGITAL ECONOMIST

Date of Publication | March 27, 2026

Expert Insights

Mickie Chandra, Nikhil Kassetty, Nithin Singh Mohan, and Melissa Tony Stires

Navigating the AI Open Seas

The Human North Star

HUMAN-CENTERED AI | ETHICS INTO ACTION | GUIDED INNOVATION

Table of Contents

Executive Summary

AI is advancing rapidly, but capability alone is not a sufficient measure of progress. This paper argues that AI development must be guided by a human North Star grounded in dignity, rights, accountability, inclusion, and well-being. Its central claim is that the future of AI should be judged not only by what systems can do but by whether they improve human lives, protect fundamental freedoms, and strengthen trust in institutions.

The paper emphasizes that responsible AI requires more than technical safety. It also requires governance models capable of addressing bias, misuse, privacy risks, exclusion, and the erosion of human agency through transparency, oversight, shared accountability, and practical mechanisms such as regulatory frameworks, institutional review, and design standards. Across organizational, societal, and individual contexts, the argument remains consistent: values must not only be articulated but translated into operational practices.

The ultimate goal is human flourishing. AI should expand opportunity, support well-being, learning, and creativity, strengthen institutional trust and democratic integrity, and serve the common good rather than narrow commercial or efficiency-driven ends.

Introduction

Artificial intelligence is often compared to an open sea: vast, full of opportunity, yet unpredictable and perilous without proper navigation. As we chart new waters with AI, a pressing question emerges: What will serve as our "Human North Star"? The rapid deployment of AI, from social media feeds to autonomous weapons, has left society searching for steady coordinates in a shifting landscape. Technological capabilities are racing ahead while ethical and governance frameworks struggle to keep pace. In the words of UN Secretary-General António Guterres, those alarmed by AI's speed "must understand a simple fact: this technology is moving exponentially" (Guterres 2023). But speed without direction can lead to disaster, much like ships in the seventeenth century that wrecked for lack of navigational reference points (Haugen 2024). More than ever, we need a guiding light rooted in human values and rights.

This white paper explores that need through six guiding questions that collectively illuminate our path forward:

- **What are our shared values?** Identifying the fundamental values that unite societies and should guide technological development.
- **How do we define inalienable rights?** Articulating the non-negotiable rights that AI must respect and protect.
- **How can we collectively stay on course toward progress and growth?** Ensuring that innovation advances societal well-being rather than diverting it off course.
- **How do we define safety, and how do we develop guardrails for the most vulnerable?** Redefining safety in the context of AI and establishing protections for individuals and communities at greatest risk.
- **What values do we hold dear as an organization, as a society, and as individuals?** Examining how values operate across institutional, societal, and personal levels and how their alignment can guide responsible AI development.
- **How do we promote the flourishing of humanity, and what will it look like?** Envisioning a future where humanity thrives alongside AI and outlining its defining characteristics.

Each section analyzes these questions in turn, drawing on insights from technologists, policymakers, educators, and thought leaders. The discussion is grounded in verified sources, ranging from speeches at the United Nations to academic perspectives and legislative actions, to avoid speculation and focus on concrete observations and recommendations. The tone is formal yet accessible, aiming to speak to a broad audience: technology professionals wrestling with ethical design, policymakers crafting governance, educators shaping the next generation's relationship with AI, and citizens navigating the digital world. Throughout, the emphasis remains on actionable insights. The goal is not only to reflect on abstract ideals but also to suggest practical ways to steer the "AI ship" toward a better horizon.

As Frances Haugen, the Facebook whistleblower, noted in a 2024 keynote, the intangible nature of today's digital economy leaves society without the traditional bearings that once guided economic life (diVittorio 2024). In a world where approximately 40 percent of economic value derives from assets that "cannot be touched or seen" (diVittorio 2024), our ancestors' methods of navigation no longer suffice. We must develop new ethical instruments and compass points. The Human North Star metaphor reminds us that our enduring values and rights can serve as fixed points of reference. By following them, societies can harness transformative technologies without losing sight of what makes us human.

The following sections explore each guiding question in detail. They examine shared values and their relevance to AI development; discuss the meaning of inalienable rights in a period of rapid technological change; and analyze how AI progress can remain aligned with human-centered objectives. The paper also considers how safety frameworks can protect vulnerable populations, how values can align across organizational, societal, and individual contexts, and what a future of genuine human flourishing alongside AI might entail. Each section integrates expert perspectives and real-world examples, from Generation Z's evolving relationship with digital life to global policy debates, to create a comprehensive picture of a practical framework for human-centered AI, guided by humanity's North Star.

1.

What Are Our Shared Values?

The first step in finding our North Star is to articulate our shared values, the common principles we collectively cherish as worthy of protection and advancement. These values function as moral coordinates, guiding how AI and other technologies should be designed, deployed, and governed. Across cultures and communities, several recurring themes emerge in discussions about technology and society. The following values consistently appear as foundational reference points.

1.1 Human Dignity and Respect

At the core of most ethical traditions lies the principle that every individual possesses inherent dignity. Timothy Shriver argues that treating people with dignity “makes life better for everyone” and has historically driven transformative social movements (Shriver 2025). Respect for persons, regardless of their background, identity, or status, is therefore a foundational value that technology should reinforce rather than erode.

Digital platforms should not dehumanize users or turn them into mere data points; rather, technology should reinforce the respect we owe one another. Shriver laments that too often, our leaders and online discourse model the opposite, contempt and dehumanization, and he calls for a return to civility and respect as baseline values in the digital age (Shriver 2025). AI systems, from customer service bots to social media algorithms, should be aligned with this respect for human dignity, avoiding practices that exploit or manipulate users.

1.2 Well-Being and Compassion

Human well-being, including mental, emotional, and spiritual health, is a universal value that technology must serve. The voices of young people illustrate this well. In a recent survey, 85 percent of Gen Z young adults agreed that their generation spends too much time on screens, and more than half said in-person relationships are more valuable than digital ones (Berry 2025). Despite unprecedented levels of connectivity, many report feelings of loneliness and fragmentation in their lives (Berry 2025).

This reveals a shared yearning for genuine human connection, balance, and mental wellness. Our technologies should be reoriented to support these needs, not undermine them. The proliferation of AI-driven content and constant notifications, for instance, should be evaluated against the values of balance, mindfulness, and compassion for the user. Shared human values would prioritize designs that promote healthy usage, protect time for family and community, and foster empathy. In short, if a technological feature or AI application is “ruining the soul” or eroding well-being, it conflicts with our core values and needs adjustment (Berry 2025).

1.3 Equity and Justice

Fairness, justice, and inclusion are also widely recognized as shared societal commitments. Societies agree that people should be treated fairly and that opportunities (and burdens) should be justly distributed. In the realm of AI, this translates into ensuring algorithms do not perpetuate bias or discrimination. Evidence shows that AI systems can inadvertently reproduce or amplify biases present in their training data; for example, by profiling individuals or marginalizing certain groups if trained on skewed data (Guterres 2023).

The shared value of equity demands vigilant guardrails so that AI systems promote equal treatment. It also means striving to make the benefits of AI accessible to all, rather than widening digital divides. When considering public policy or organizational use of AI, one must ask: Does this align with our societal value of justice? Does it help all communities, or only a privileged few? Recognizing equity as a North Star value guides us to design interventions like bias audits, diverse training data, and inclusive user research to uphold fairness.

1.4 Accountability and Transparency

Across the board, there is a call for accountability in the tech industry, a value that has been lacking and must be reinforced. Frances Haugen highlighted that most big tech companies lack a “guiding force” of transparency or accountability, which led to decisions favoring profit over people (Haugen 2024). The shared human value here is honesty and accountability: we expect those who wield power (including the power of advanced AI) to be answerable for impacts on society. Transparency is a part of this value set, openness about how AI systems make decisions, what data they use, and what outcomes they drive. The public increasingly demands this openness, as seen in multiple jurisdictions passing data privacy and AI ethics laws (diVittorio 2024). Embracing transparency as a core value can help rebuild trust.

For example, Haugen notes the absence of meaningful transparency metrics in social media platforms, unlike car companies that are held to safety tests, and suggests that companies should adopt clear metrics for societal impact (Claburn 2024). A shared commitment to truth and accountability can thus direct organizations to implement audit trails for AI, publish impact assessments, and welcome external oversight.

1.5 Community and Solidarity

Humans are social creatures, and a value that emerges strongly is our commitment to community: caring for one another and the common good. In technological terms, this implies that AI should strengthen communities rather than isolate individuals. It also means we value collaboration and collective action. The challenges posed by AI (from job disruption to ethical dilemmas) are not ones any one person or nation can solve alone. They require coordinated responses among governments, industry, academia, and civil society. This emphasis on collaboration reflects a deeper value of solidarity. We see this in calls for international cooperation on AI governance and in multi-stakeholder initiatives to create ethical standards. The underlying value is that we are "all in the same boat," and thus, we have a shared responsibility to each other.

This sensibility harkens back to Shriver's observation that in the past, people "risked everything for freedom and for each other and for a future many did not live to see" (Shriver 2025). Such solidarity, whether in world war or civil rights struggles, was fueled by values of brotherhood, altruism, and hope for a better collective future. As we navigate the AI age, rekindling this spirit of common purpose is vital. It reminds us that AI's purpose is not to pit us against each other in competition but to help us solve shared problems like disease, climate change, and poverty.

1.6 Dario Amodei on Shared Values and Empathy in AI

Dario Amodei, CEO of Anthropic, argues in his essay “Machines of Loving Grace” that AI should be designed to reflect and reinforce human values, particularly emphasizing empathy and the alignment of AI systems with societal well-being. Amodei stresses that AI technologies should not only be responsible in their technical design but also deeply embedded with empathy for human users. This means ensuring that AI systems understand human intentions, respect privacy, and promote fairness, safeguarding against misuse. By instilling these values into AI systems from their design phase, Amodei envisions a future where AI not only serves the economic interests of society but also helps advance the common good. This approach directly connects to the shared values of human dignity, well-being, and justice, reinforcing the idea that technology should prioritize human-centric outcomes and align with the needs of society as a whole (Amodei 2025).

Taken together, these shared values can be thought of as the human constants by which we steer technological change. These include respecting the intrinsic dignity of each person, promoting overall well-being and compassion, ensuring fairness and justice, insisting on accountability for those who create powerful systems, and working together in a spirit of community. These values are not new; they are ancient and enduring. But as we integrate AI into the fabric of daily life, these values must be consciously reasserted and woven into the design and deployment of technology. They form the criteria by which we judge whether AI is on the right course. When a new AI application is introduced, we should ask: Does this uphold or undermine human dignity? How does it impact well-being and mental health? Is it fair to all groups? Is it developed and run transparently? Does it bring people together or drive them apart? By filtering our innovations through these value questions, we align them with our North Star.

Crucially, identifying shared values is only the start. We then face the hard work of implementing them. As Haugen put it, without a North Star, “acceptability can shift over time,” what we deem normal or permissible might drift away from our true values (Haugen 2024). Having established what we stand for, society must hold the line and integrate those standards into AI governance. In the next sections, we build on these values by examining the concrete rights they imply, and how to enforce those rights and values through practical guardrails and collective efforts.

2.

How Do We Define Inalienable Rights?

If shared values are our guiding stars, inalienable rights are the fixed constellations in our moral sky, enduring, non-negotiable entitlements that every human being holds. These rights are “inalienable” because they cannot be given up or taken away. They are often enshrined in constitutions, international law (like the Universal Declaration of Human Rights), and ethical traditions worldwide. The challenge in the context of AI is to define and enforce these rights in new settings. As technology permeates everyday life, we must ask: Which human rights are most at stake, and how do we safeguard them in practice?

Traditionally recognized inalienable rights include the right to life, liberty, and personal security; freedom of thought and expression; privacy; equality and non-discrimination; and the right to an adequate standard of living, among others. All these can be impacted (positively or negatively) by AI. For example, AI in healthcare could enhance the right to health by enabling medical breakthroughs, but AI-driven surveillance could undermine the right to privacy or free assembly if misused by authoritarian regimes. Therefore, defining inalienable rights in the age of AI involves reaffirming classic rights and extending them to digital contexts.

2.1 Right to Privacy

Privacy is widely regarded as a fundamental right, the right of individuals to control their personal information and keep certain aspects of life free from intrusion. AI and big data analytics pose significant challenges in this regard, as they enable unprecedented collection and analysis of personal data. When AI algorithms harvest online behavior, or when facial recognition cameras monitor public spaces, the right to privacy is at risk.

Defending privacy in the AI era means instituting strict data protections, transparency about data use, and giving individuals meaningful agency over their own data. Legislative efforts, such as data protection laws and AI-specific regulations, aim to codify these protections. We see growing momentum in this direction: more than twenty US states have passed comprehensive data privacy laws in recent years to grapple with these issues (diVittorio 2024).

Privacy must be treated as inalienable, not something one automatically forfeited by participating in digital services. Even as AI systems rely heavily on data, they can and should be designed to anonymize or protect personal details by default. The principle of privacy by design embodies this approach, treating privacy as a right rather than a privilege.

2.2 Right to Freedom from Bias and Discrimination

In human rights terms, this corresponds to the right to equality and freedom from discrimination based on race, gender, religion, or other protected characteristics. AI systems must be held to this standard as well. Because AI models can inherit biases from historical data or biased design choices, there is a real risk that automated decisions (whether in hiring, lending, policing, etc.) could unfairly disadvantage certain groups.

From a rights perspective, freedom from algorithmic bias can be understood as an extension of established anti-discrimination protections. For instance, António Guterres has warned that AI "can amplify bias [and] reinforce discrimination" if appropriate safeguards are not implemented (Guterres 2023).

An inalienable rights approach would require that individuals have recourse if an AI system makes a prejudiced decision affecting them, essentially treating algorithmic fairness as a rights-based issue. Some scholars and policymakers have proposed a "right to explanation" in AI, meaning that individuals adversely affected by automated decisions should have the right to understand how those decisions were made and to challenge them. This principle connects directly to due process rights and equality before the law, updated for the algorithmic age.

2.3 Right to Safety and Security of Person

Everyone has the right to be safe and protected from harm. In an AI context, this right extends to both physical and psychological harms that AI technologies may facilitate.

Physical safety concerns include issues like autonomous weapons (often called "killer robots") that Guterres has argued should be banned on both moral and security grounds (UNA-UK 2023). There is growing international concern that lethal autonomous weapons systems, which can make life-and-death decisions without meaningful human intervention, violate the right to life and undermine the principle of human control in warfare.

Similarly, AI-driven disinformation and deepfakes can endanger people's security by inciting violence, damaging reputations, or enabling harassment. Legislative responses are beginning to reflect this reality. California's recent AI-related legislation, for example, explicitly criminalizes AI-generated child sexual abuse material and non-consensual sexual deepfake images as crimes (Wahab 2024). By extending these laws to cover synthetic media, the legislation recognizes the right of individuals (especially women and children) to not be exploited or victimized by these AI tools.

State Senator Aisha Wahab, who authored the legislation, emphasized that the intent is to "protect vulnerable populations on the internet" through necessary guardrails (Wahab 2024). Here, we see the concept of safety as a right, the idea that people have a right to digital safety just as much as physical safety.

2.4 Right to Information and Transparency

In a world increasingly mediated by AI systems, a compelling argument can be made that people have a right to know when AI is being used and how it affects them. This principle can be framed as part of the broader right to information, or even as an extension of freedom of thought.

If AI systems curate the news we read or shape the social media content we encounter, should individuals have the right to understand how those algorithms function? Advocacy from Frances Haugen has emphasized the issue, calling for greater transparency in understanding how online platforms rank and moderate content (Claburn 2024).

Some jurisdictions are now considering laws that require labeling of AI-generated content (to protect the right not to be misled) or disclosure when chatbots, not humans, are interacting with customers. These moves reflect an emerging value: people's right to know and not be deceived by AI systems. It complements classical rights like freedom of expression, which is undermined if automated systems covertly distort the public discourse.

2.5 Right to Human Agency and Identity

There is also a deeper philosophical right at stake: the right to remain fundamentally human in the face of technologies that could alter or usurp human agency. John Kennedy Philip, in his history of transhumanism, notes that proponents of human enhancement seek to overcome human limitations (Philip 2024). While extending life or intellect with technology could be seen as an expression of individual liberty, critics like Francis Fukuyama fear that such transhumanist projects could "violate basic human rights" and blur the line of what it means to be human (Philip 2024). This raises the question: Do we have a right to preserve our human identity and agency against extreme technological alteration? If technologies one day allowed mind uploading or genetic engineering of "designer babies," societies would need to determine how to balance innovation with the rights of individuals (and even future generations) to an open future and an uncoerced, natural development. These are complex ethical issues, but the principle is that certain experiments might cross ethical red lines.

Defining inalienable rights in the AI era may therefore include protecting core aspects of human agency: the ability to make meaningful choices, to retain one's humanity rather than being reduced to data, and to ensure that technology augments rather than replaces human decision-making. UNESCO's Recommendation on the Ethics of AI (2021), though not one of our specific sources, is an example of an international effort to articulate such principles, including respect for human autonomy.

In defining inalienable rights today, we are not writing on a blank slate. We have a rich legacy of human rights laws and norms. The task is to interpret and enforce them in light of AI's new challenges. Klaus Stoll and Sam Lanfranco (2023) argue that we may need an "AI-specific Human Rights...Trustmark" to certify that technologies meet human rights standards. Their proposal, HARIET (Human Rights AI Evaluation Trustmark), is essentially a tool to ensure AI systems are audited for human rights compliance. The idea is that a reliable indicator can help consumers and institutions identify which AI products uphold fundamental rights (Stoll and Lanfranco 2023). They describe HARIET as providing "human oversight and ethical guidelines based on the fundamental Human Rights that are common to all humanity to act as guardrails against AI abuse" (Stoll and Lanfranco 2023). This is a concrete approach: rather than just declaring rights in principle, incorporate them into certification and design processes.

To illustrate how rights definitions translate into concrete policy, consider again the California legislation discussed earlier (Wahab 2024). It effectively operationalizes several rights simultaneously: the right of a person not to have their likeness used in fake pornography without consent (a mix of privacy, dignity, and safety rights), and the right of minors not to be depicted in abusive sexual content (protecting children's rights and safety). By updating existing laws (like those on harassment and child exploitation) to include AI-generated media, the law asserts that digital harms are real harms against one's rights. This reflects a broader understanding that our rights to safety, reputation, and psychological integrity must extend into cyberspace.

In summary, inalienable rights in the AI era encompass both enduring human rights and newly salient digital rights. We define them as follows: everyone has the right to privacy and control over personal data; the right to be free from algorithmic discrimination; the right to safety and security in digital as well as physical environments; the right to transparency and truthful information about AI interactions; and the right to exercise human agency, with technology enhancing rather than degrading our humanity. These rights are non-negotiable. They set the boundaries within which AI innovation can occur. If an AI application threatens to cross these boundaries, it must be reined in by policy or design changes.

It is important to acknowledge that rights can sometimes conflict. For instance, one person's right to free expression might conflict with another's right to be free from hate speech. Navigating those tensions in the context of AI will require thoughtful policy and perhaps new jurisprudence. But having a clear definition of priority rights provides a framework for adjudication. For example, if an AI-driven content platform is spreading harmful misinformation, we weigh the right of the public not to be harmed or deceived against abstract claims of the platform's freedom to deploy any algorithm. The balance should favor protecting fundamental human rights.

Ultimately, defining inalienable rights in relation to AI sets the stage for our next question: How do we stay on course toward progress and growth while respecting these rights and values? The recognition of rights draws lines in the sand. It tells us what not to do (e.g., do not violate privacy, do not allow AI to be racist, do not let people come to harm through AI negligence). The next section examines the proactive side: Given these moral boundaries, how do we chart a positive course that harnesses AI for human progress?

3.

How Can We Collectively Stay on Course Toward Progress and Growth?

Navigating the “AI open seas” is not a solo journey. It requires collective effort. The question of staying on course toward progress and growth asks: How do we ensure that, as a society, we steer AI in a direction that fosters positive development rather than running aground on ethical, social, or economic shoals? Progress and growth here refer not only to economic advancement but also to holistic human progress: improvements in quality of life, knowledge, sustainability, and societal well-being. Achieving this means avoiding the hazards (bias, inequality, misuse) while capturing AI’s benefits for all. Several key strategies and principles emerge from expert insights on how to keep our course true.

3.1 Establish a Clear “North Star” Vision and Metrics

We cannot stay on course if we have not agreed on where we are going. Frances Haugen offers an instructive maritime metaphor: She noted that in the past, ships literally drifted because they lacked “absolute references” for navigation (diVittorio 2024). In today’s context, our North Star must be a clear set of ethical goals and metrics for AI aligned with human values.

Haugen (2024) warns that “without a north star, acceptability can shift over time,” companies may gradually normalize practices that sacrifice user well-being if profit becomes the only metric. To counter this, organizations and governments should explicitly define what responsible AI looks like and measure progress against it.

For example, a social media company might adopt a North Star goal of “maximizing meaningful social connection” rather than just maximizing engagement time. It could then track metrics such as user well-being or the quality of civic discourse rather than focusing exclusively on advertising clicks. Collectively, industries might also establish targets for reducing AI-driven harms (such as a goal to cut the prevalence of algorithmic discrimination or misinformation by a certain percentage).

By developing these ethical performance indicators, we create a feedback loop that helps correct course when we drift. As Haugen observed at the DataGrail Summit, new “methods for measuring success” are needed in the intangible digital economy, along with collaboration to establish the “boundaries of what’s just and what’s not” (diVittorio 2024). In practice, this could take the form of industry standards, ethical AI certifications, or government guidelines that clearly articulate the direction of progress.

3.2 Embrace Multistakeholder Collaboration and Global Governance

AI is a global phenomenon; its development and impacts do not stop at national borders. Thus, staying on course is something any single actor can accomplish alone. It requires collective governance.

António Guterres (2023) has emphasized that while many countries are proposing different AI initiatives, "this requires a universal approach" because AI systems operate across jurisdictions. One concrete proposal he has raised is the creation of a new UN entity dedicated to AI governance, analogous to agencies like the International Atomic Energy Agency, to coordinate global efforts (Guterres 2023).

The rationale is that, just as nuclear technology required an international governance framework to ensure it contributed to progress (e.g., nuclear energy) while minimizing risks (e.g., nuclear proliferation), AI may also require a global body that sets norms and shares best practices.

Moreover, multistakeholder collaboration, involving not just governments but also tech companies, civil society, academia, and affected communities, is crucial. Each stakeholder brings a different perspective and expertise: companies understand the tech, civil society voices ethical and social concerns, researchers contribute knowledge, and citizens/users highlight lived experiences.

Forums for such collaboration (like standards committees, summits, or multi-sector advisory councils) help keep the development of AI aligned with widely shared interests rather than narrow agendas. The Summit of the Future, scheduled by the UN, along with various AI safety conferences, are examples of venues to align global visions.

Collectively staying on course also means building consensus on guidelines (for instance, agreeing that AI should not be used for social scoring that violates human rights, or agreeing on limits for autonomous weapons). It also requires capacity building, helping less-developed nations participate in AI's benefits so that progress is inclusive rather than one-sided.

The overall point is that progress guided by one country or company alone can easily veer off track; progress guided by the international community has a better chance of reflecting humanity's broader good.

3.3 Prioritize Ethical Innovation Over Profit-Only Motives

The engine of AI progress has largely been private sector innovation, often driven by competition and profit. While market incentives have spurred rapid advances, they can also cause innovation to deviate from the public interest. To stay true to a course of broad human progress, the innovation ecosystem must internalize ethical considerations, not just financial ones. Frances Haugen's experience at Facebook illustrated that without a guiding principle favoring people, companies made trade-offs "time and again that favored profit rather than people" (Haugen 2024).

Collectively, we can address this by reshaping incentives. Governments can implement regulations that penalize harmful AI outcomes and reward safety and fairness (for example, tax incentives for developing AI for accessibility or grants supporting AI-for-social-good projects). Investors can also play a role by adopting ESG (environmental, social, governance) criteria that evaluate how tech companies manage AI ethics, thereby encouraging firms to value long-term societal impact.

Within organizations, leaders can cultivate cultures where ethical concerns are raised and addressed during product development. One actionable idea is requiring an "ethical impact assessment" alongside every major AI deployment, similar to environmental impact assessments used in development projects.

This shifts the mindset from "Can we build it?" to also asking "Should we build it, and how do we build it right?" By making ethical innovation the norm, the collective path of progress becomes steadier. It means breakthrough AI in fields like medicine or education is celebrated and pursued, whereas AI applications that purely exploit human weaknesses (like excessively addictive algorithms) are viewed skeptically. As Haugen put it, it's about getting to a place "where we're not afraid to innovate" but have the tools to maximize positive impact (Claburn 2024). Responsible innovation frameworks are such tools.

3.4 Align AI Progress with Sustainable Development Goals

One way to define a positive course for AI is to explicitly tie it to solving humanity's greatest challenges, many of which are outlined in the United Nations Sustainable Development Goals (SDGs). Guterres passionately calls for a "race to develop AI for good," AI that can help end poverty, banish hunger, cure diseases, and combat climate change, effectively propelling us toward the SDGs (Guterres 2023).

If AI advances are evaluated by how much they contribute to these goals, attention naturally shifts toward socially beneficial innovation. For instance, AI progress in agriculture should be measured by how effectively it improves crop yields or food distribution for the hungry, not just by its novelty. Similarly, AI progress in climate science should be evaluated by contributions to renewable energy optimization or climate prediction models that help vulnerable communities mitigate environmental risks.

By orienting research funding, competitions, and recognition toward AI-for-good initiatives, we collectively push the ship in the right direction. There are promising signs: many governments and institutions are funding AI projects in healthcare (like AI for faster vaccine development), in education (AI tutors for underserved areas), and in environmental protection (AI for wildlife conservation). However, we also see enormous talent and resource pools being used for less lofty aims (e.g., optimizing ad clicks or speculative financial trading algorithms).

A recalibration may therefore be necessary. Public-private partnerships could help redirect AI expertise toward public interest projects. The collective benefit is twofold: we achieve progress on real human problems, and we maintain public support for AI because people see tangible good being created.

3.5 Develop Adaptive Governance, Learn and Course-Correct

Staying on course is not a one-time decision; it requires continuous monitoring and the ability to adjust as circumstances evolve. The metaphor of a ship's journey is apt. Even with a fixed destination, sailors must constantly adjust to winds and currents. Similarly, governments and organizations should practice adaptive governance for AI. This means regularly reviewing outcomes, sharing lessons across borders, and updating rules and best practices accordingly.

For example, if a certain AI application (like a predictive policing tool) is deployed and later evidence shows harmful effects, mechanisms must exist to rectify or halt its use. Adaptive governance must include sunset clauses for high-risk AI systems, requiring re-evaluation after a defined trial period. It could also include stronger feedback channels through which researchers, civil society groups, and affected communities inform policy decisions.

Multistakeholder advisory boards, such as the High-Level Advisory Board for AI proposed by Guterres, can also play a role by evaluating governance options and recommending mid-course corrections (Guterres 2023).

The dynamic nature of AI (e.g., the sudden emergence of generative AI like ChatGPT) means our governance must also be dynamic. The collective aim is to stay proactive rather than reactive. By identifying early warning signs (say, an uptick in AI-related job displacement or a new form of deepfake fraud), we can take coordinated actions (like workforce retraining programs or fraud detection initiatives) to keep progress on a socially beneficial track.

3.6 Foster Public Engagement and Shared Accountability

Finally, a democratic and inclusive process is key to staying on course. The public should have a say in the direction of AI progress, not just experts. This shared accountability means educating citizens about AI (so they can participate meaningfully) and inviting public input on major AI policies. When people feel empowered and informed about AI's trajectory, they can help correct course through consumer choices, voting, and advocacy.

Public opinion can act as a compass, a means of collective course correction if technologies drift away from societal values. For instance, strong public concern about privacy can pressure companies to prioritize privacy-preserving technologies.

Timothy Shriver's work on the Dignity Index, where citizens evaluate language in public discourse, illustrates how public engagement can steer cultural norms toward dignity (Shriver 2025).

A similar approach could apply to AI governance: citizens' juries, public consultations, or town halls discussing the use of technologies such as facial recognition in policing and guiding officials on what is acceptable. Such grassroots engagement ensures that "progress" remains aligned with how people actually define progress in their lives (safety, happiness, opportunity, etc.) rather than a purely technocratic or profit-centric definition.

In sum, collectively staying on course toward progress and growth with AI involves setting a clear values-driven vision, cooperating globally, tweaking our innovation incentives, aiming AI at humanity's grand challenges, being ready to adapt governance as we learn, and involving the public in decision-making.

It is a tall order, but we have historical precedents to draw on. Humanity navigated the advent of powerful technologies before (printing press, electricity, atomic energy) by eventually creating norms and institutions to guide them. The key is ensuring that our proverbial compass, consisting of ethical principles and human rights, remains properly calibrated and consistently consulted.

We should also acknowledge that "progress and growth" can mean different things to different communities. By engaging in broad dialogue, exactly the kind of multi-perspective analysis in this paper, we can refine a collective understanding of progress measured in human-centric outcomes.

Growth is not merely GDP expansion or tech-sector valuation; it's growth in human flourishing, knowledge, and capability. Keeping that holistic definition front and center helps prevent us from being led astray by narrow metrics.

Frances Haugen offers an optimistic note: despite the uncertainties of the "intangible economy," humanity has "innovated our way to clarity" during past periods of uncertainty (diVittorio 2024). We devised longitude measurements after many perilous voyages. By working together now, "we can stay afloat, and even sail, if we work together," and equip ourselves with the right tools to make the most impact (Haugen 2024). In concrete terms, those tools are ethical frameworks, cooperative institutions, and an engaged society. With them, collective progress not only becomes possible; it becomes our likely destiny.

4.

How Do We Define Safety? How Do We Develop Guardrails for the Most Vulnerable?

As we pursue progress with AI, we must also ask: What does “safety” mean in this new context? Traditionally, safety might refer to physical safety (preventing accidents, harm, or death). In the AI context, safety has a broader meaning, encompassing not only the physical integrity of people but also the protection of mental health, privacy, democratic processes, and social stability from AI-related harms. Furthermore, the question specifically highlights the most vulnerable: those who are at greatest risk of being harmed or left behind by unchecked technological change. Defining safety, then, involves identifying the risks AI poses and determining what safeguards (“guardrails”) can mitigate those risks, with special attention to protecting vulnerable populations.

4.1 Redefining “Safety” in the AI Era

Safety in AI can be viewed along multiple dimensions:

- **Technical Safety:** Ensuring AI systems reliably do what they are intended to do, without dangerous failures. This includes preventing self-driving cars from causing crashes or AI systems from behaving erratically in unintended ways. It also covers cybersecurity, making sure AI cannot be easily hacked and turned malicious. Technical safety is akin to the engineering challenge of building robust systems.

- **Ethical and Societal Safety:** Ensuring AI does not cause harm through its decisions and broader impacts on people's lives and society. This includes preventing harm such as biased judgments (which may damage someone's opportunities), harmful content dissemination (such as AI-generated hate speech or extremist propaganda), invasion of privacy, or enabling authoritarian oppression (like mass surveillance or social credit systems). It also includes psychological safety, preventing harms such as online harassment or mental distress amplified by AI-driven social media algorithms.
- **Long-Term Existential Safety:** A less immediate but often discussed aspect, which ponders the risk of highly advanced AI becoming uncontrollable or harmful to humanity as a whole (the "AI alignment" problem). While speculative, some experts include this in definitions of AI safety.

For the purposes of this discussion, and given our sources, the focus is on ethical and societal safety, guarding against current, tangible harms, especially to vulnerable groups. Now who is "the most vulnerable"? In context, this often means children, marginalized communities, individuals lacking digital literacy, or those who could be disproportionately harmed by AI decisions (e.g., people subject to profiling or workers whose jobs may be displaced without adequate support). It can also mean society at large when democratic systems or public security become vulnerable to AI-enabled threats.

Let us examine specific areas of concern and the corresponding guardrails being developed or proposed.

4.1.1 Protecting Children and Minors Online

One clear area of vulnerability is children exposed to AI-powered platforms. Children are susceptible to exploitation, cyberbullying, or exposure to harmful content, and they cannot always discern manipulation. AI can generate deeply inappropriate content (such as deepfakes or explicit imagery) that puts children at risk.

Policymakers have increasingly treated this as a critical safety issue. In California, Senator Aisha Wahab's AI Child Safety Act (SB 1381) was signed into law, targeting exactly these problems (Wahab 2024). Part of her legislative package also includes the Stop the Online Predators Act (SB 926) and the Digital Identity Theft Act (SB 981), each addressing facets of AI-enabled exploitation.

These laws expanded legal definitions of child pornography to cover AI-generated images and made it illegal to create or distribute sexually explicit deepfakes without consent, particularly in revenge pornography scenarios (Wahab 2024). For example, Wahab noted that up to 95 percent of deepfake videos since 2018 have been non-consensual porn, and 90 percent of those victims are women (Wahab 2024). This staggering statistic underscores the scale of harm enabled by such technologies.

Guardrails implemented through these laws include requiring social media platforms to provide reporting mechanisms allowing individuals to report synthetic explicit images posted without consent (Wahab 2024). It also means perpetrators can be prosecuted under existing laws once AI-generated material is explicitly covered. These measures strengthen safety by closing legal loopholes, making it clear that abusing AI tech to harm someone's dignity or safety is unlawful.

Beyond legislation, protecting minors may also involve technical guardrails: age-appropriate AI systems (like YouTube's restricted modes or AI tutors with content filters), parental controls that leverage AI to detect risky interactions, and robust content moderation using AI to flag grooming behavior or extreme content aimed at children.

Safety-by-design principles suggest that any platform likely to host minors should include built-in protections (for example, not recommending self-harm content, limiting data collection on kids, etc.). The overarching idea is that those least able to protect themselves, children, should get extra layers of protection through both law and technology.

4.1.2 Combating Harassment, Exploitation, and Mental Harms

Vulnerable groups also include women, minorities, and individuals who may become targets of harassment or exploitation online. Social media platforms driven by AI algorithms can sometimes intensify such harms by amplifying outrage or enabling anonymity for aggressors. Guardrails in this context may include algorithmic adjustments and moderation systems designed to detect and reduce harassment.

One important guardrail is transparency. If companies disclose how their algorithms prioritize content, external researchers and watchdog organizations can identify designs that produce harmful outcomes. Frances Haugen pointed out that currently, “most social media platforms have no transparency metrics” for comparing safety performance (Claburn 2024). Her advocacy implies that transparency and independent audits could push platforms to fix dangerous algorithmic tendencies, such as prioritizing sensational or divisive content that may increase harassment or social instability.

Mental health concerns are also significant. Heavy social media usage and constant information flows have contributed to what Taylor Berry describes as a “fragmentation” of attention and increasing loneliness among the younger generation, with one in three Gen Z young adults reporting persistent loneliness (Berry 2025).

From this perspective, young people themselves may be considered a vulnerable group in terms of mental well-being. A safety-oriented AI paradigm might therefore include digital well-being features: for example, systems that encourage healthier usage patterns by nudging users to take breaks or guiding them toward supportive communities rather than harmful content loops.

Some platforms have already begun integrating such features, including screen-time monitoring tools and content warnings. These serve as soft guardrails, helping reduce psychological harm. The values of well-being and safety overlap here; protecting users from addictive or manipulative patterns is part of keeping them safe in a holistic sense.

4.1.3 Ensuring Non-Discrimination and Justice

Marginalized communities, including ethnic minorities and the economically disadvantaged groups, can be especially vulnerable to biased AI outcomes. A classic example: an AI used in criminal justice might unfairly predict higher recidivism risk for people of color due to biased historical data, leading to harsher sentences or bail decisions, thus violating their rights and safety (freedom) unfairly.

Guardrails against these risks include algorithmic bias testing, fairness standards, and mandatory human review for high-stakes decisions. There is a growing momentum behind "algorithmic accountability" laws that require impact assessments when AI systems are deployed in sensitive domains such as employment, lending, or policing. The idea is to catch and correct biases before they cause harm.

Another guardrail is the principle of human oversight. Crucial decisions affecting individuals' lives, such as hiring, medical diagnoses, and legal decisions, should not be left solely to automated systems. Human judgment must remain involved to ensure context and compassion are considered, especially when a person's well-being is on the line.

Stoll and Lanfranco's HARIET trustmark concept ties in here as well. Their framework proposes that one function of HARIET would be to provide ethical guidelines based on fundamental human rights as guardrails against AI abuse (Stoll and Lanfranco 2023). For instance, HARIET would flag an AI system that lacks appeal mechanisms or that processes personal data in ways that could violate privacy rights. By certifying AI that meets human-rights standards, it indirectly pressures creators to build safer systems.

4.1.4 Guarding Against Misinformation and Threats to Democracy

The health of our information environment is a societal vulnerability. AI can generate highly convincing fake news or highly realistic deepfake videos capable of misleading large audiences. Groups with limited media literacy—or communities exposed to targeted propaganda—may be particularly vulnerable to these tactics. Safety in this domain means protecting the integrity of public discourse and democratic institutions. Potential guardrails include authentication systems for media content (for example, watermarking AI-generated material so it can be identified) and automated fact-checking systems that rapidly detect and debunk false claims.

Governments and civil society groups have begun developing frameworks to address these threats; for example, the EU's initiatives to have AI platforms assess the risk of disinformation and report how they handle it. While our sources don't directly cover an example of this, it's implicitly related to the calls for transparency and ethics in AI.

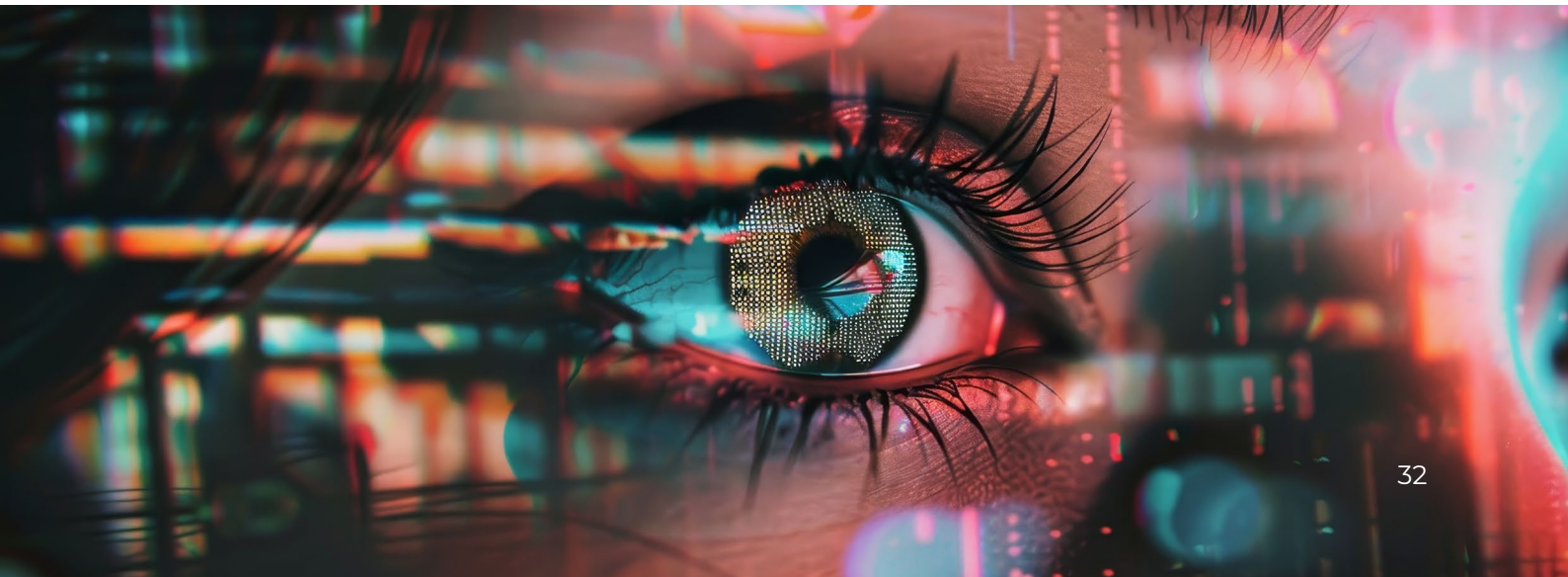
Haugen's metaphor about safety rituals needing to change is relevant. She noted industries such as automobile manufacturing use independent crash tests to enforce safety standards, but with opaque algorithms, equivalent safety rituals (Claburn 2024) are often absent.

Adapting that insight, one could argue we need the equivalent of crash tests for algorithms, independent evaluations to see how an AI could "crash" society via misinformation or manipulation, and then require fixes before deployment.

4.1.5 Focusing on Those Left Behind, Economic and Educational Vulnerabilities

Another aspect of vulnerability is those who might be displaced or disadvantaged by AI economically. For example, low-skilled workers might lose jobs to automation; students in poorer school districts might not have access to beneficial AI educational tools that others do. Safety, in a broad sense, therefore, includes economic security and not exacerbating inequality.

Guardrails in this area might be policy measures like reskilling programs funded by tech gains, or investing in equitable AI in education so that underserved communities get AI tutors or personalized learning to catch up (Lim, Hilmy, and Wei 2025). The UNESCO piece by Lim and colleagues envisions an education future where students are empowered rather than made passive by AI. The guardrail implicit there is ensuring the AI doesn't deskill students or worsen gaps. Instead, it should "serve rather than shape educational goals" (Lim, Hilmy, and Wei 2025). That is a powerful guiding guardrail: technology must serve human goals instead of distorting them.



If we see AI leading to negative outcomes like a generation of students who can't think critically because they over-relied on AI, that means safety in education was compromised. To guard against it, educators can incorporate AI literacy and "metacognition support," teaching students how to use AI as a tool without losing their own agency (Lim, Hilmy, and Wei 2025).

In broader economic terms, guardrails might also include strengthened social safety nets or policy discussions around radical proposals such as universal basic income if AI upheaval becomes extreme. These are socioeconomic guardrails complementing technical ones.

4.2 Implementing Guardrails: The How

Recognizing the need for guardrails is one thing; building them is another. We have touched on legislation and corporate policy changes, but it's worth summarizing key approaches:

- **Laws and Regulations:** Clear legal boundaries (such as the California AI bills) define unacceptable behavior and provide enforcement. Regulations can also mandate standards (for example, requiring AI systems in certain domains to undergo bias audits or certification). Senator Wahab's statement encapsulates this approach: "We must ensure we are...requiring tech companies to add necessary guardrails that protect vulnerable populations" (Wahab 2024). Law turns moral imperatives into compliance requirements.
- **Industry Self-Regulation and Best Practices:** Sometimes, industries create ethical guidelines (like AI ethics boards or partnerships) to police themselves. While not a substitute for law, these can produce quick, flexible standards. For instance, some tech companies voluntarily ban the use of their facial recognition tech by law enforcement pending better rules, acknowledging potential harm to minority rights. This is a kind of self-imposed guardrail.

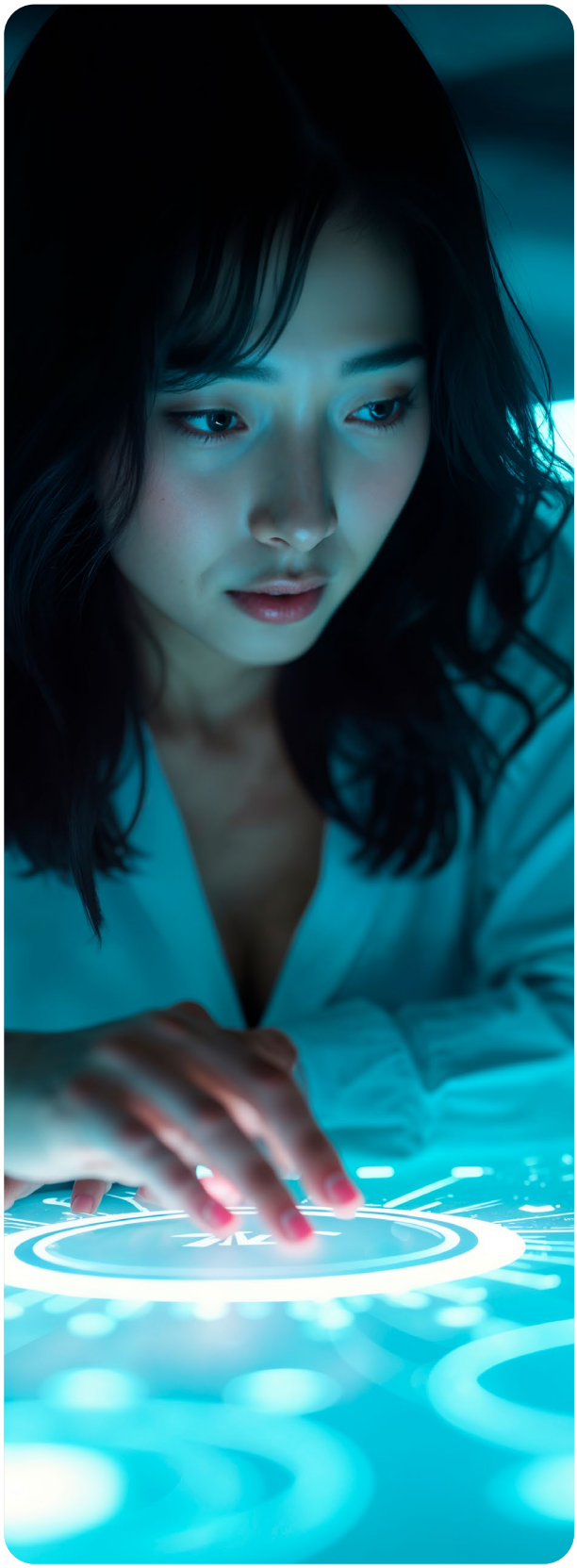
- **Technology Design Solutions:** Embedding safety in design; for example, building AI that can explain its decisions (to facilitate human oversight) or that has fail-safes if it detects it's operating outside its parameters. AI developers working on advanced models are also researching how to make models less likely to produce dangerous outputs (like instructions for illicit activities). OpenAI and others have put some guardrails on their public models (with mixed success), such as content filters. We can anticipate continued improvements as society demands safer AI by default.
- **Education and User Empowerment:** Teaching users, especially vulnerable ones, how to navigate AI critically is itself a guardrail. Digital literacy programs, warnings (similar to public health warnings but for things like deepfakes or scam AI calls), and tools that empower users to set their own safety preferences (like robust filters or time limits) all contribute.

Finally, focusing on the most vulnerable reflects a longstanding human-rights principle: the test of a society is measured by how it treats its most vulnerable members. Apply this to AI: if our AI-driven systems work well for children, for minorities, for those with disabilities, etc., then they are likely safe and beneficial for everyone.

Conversely, if we see these groups being harmed, that's an early warning that our course is wrong. Klaus Stoll and Sam Lanfranco phrase it as making human rights matter in AI, which inherently is about protecting those at risk, and they argue that human rights can be a business model, not an obstacle (Stoll and Lanfranco 2023). By reframing safety and ethics as integral to the innovation, we ensure the journey doesn't jettison the vulnerable overboard.

One poignant example to close this section: after describing the rash of AI-generated explicit images causing real trauma, Wahab noted how such abuse “can ruin an individual’s entire life,” leading to mental health deterioration and isolation (Wahab 2024). If we keep those stakes in mind, entire lives can be upended by AI misuse. It’s clear why robust guardrails are non-negotiable. Safety ultimately means preserving the integrity, freedom, and well-being of human lives in the presence of powerful technologies. Guardrails, whether legal, technical, or social, are the price we willingly pay to ensure technology remains a servant to humanity, not the other way around.

How organizations practically balance speed-to-market with ethical compliance is addressed further in the sections “How Can We Collectively Stay on Course Toward Progress and Growth and “What Values Do We Hold Dear as an Organization, as a Society, and as an Individual,” where frameworks such as ethical impact assessments and adaptive governance offer concrete guidance.



5.

What Values Do We Hold Dear as an Organization, as a Society, and as an Individual?

Values guide behavior at every level of human organization. However, the values we emphasize can differ when we talk about an individual person, a society or culture, or a formal organization (like a company or institution). Ideally, these layers of values should reinforce each other. Individuals bring personal ethics into organizations, organizations reflect societal values in their missions, and society, in turn, is shaped by the values of its members and institutions.

In the context of the AI revolution, examining values at each level is crucial because misalignment can cause problems. For example, a tech company might value rapid growth (an organizational value), whereas society values safety and individuals value privacy; if these conflict, it leads to trouble. Let's break down the values at each level and explore how they relate to our navigation of AI's open seas.

5.1 Organizational Values (Companies, Governments, Schools)

Organizations are structured groups with specific goals, and they adopt values (sometimes explicitly in mission statements, other times implicitly through culture) that drive decisions. In the tech industry, many organizations have historically prioritized efficiency, innovation, and profit. These are not inherently problematic—innovation is positive and financial sustainability is necessary—but if they are pursued to the exclusion of human-centric values, problems arise.

Frances Haugen's insider view of Facebook revealed that because there was "no North Star for what transparency looked like," the company repeatedly made choices favoring profit over people's well-being (Haugen 2024). This indicates a value gap at the organizational level. To correct that course, organizations, especially those developing or deploying AI, need to hold certain values dear:

- **Accountability and Transparency:** An organization should value being accountable to its users, customers, and the public. This means owning up to mistakes, being open about operations, and inviting oversight. Thomas Claburn notes that Haugen criticized big tech companies for the fact that "the only expectation we have on transparency and clarity is around profit and loss numbers," whereas metrics for societal impact are lacking (Claburn 2024). If organizations held transparency as a core value, they would proactively share information about how their AI systems work and affect stakeholders. An accountable organization might, for instance, publish regular reports on the societal outcomes of its AI (such as content moderation effectiveness or AI's impact on user mental health), not just its quarterly earnings.
- **Ethical Innovation and "Do No Harm":** Organizations in the AI space should adopt an ethic akin to the medical "do no harm" principle. This means valuing user safety, privacy, and fairness from the outset. It's the difference between asking "Can we deploy this feature?" and "Should we deploy this feature, and how will it affect people?" Many tech companies are now forming internal ethics teams to embody this value, but these efforts must be backed by leadership commitment. A company that truly values ethics will empower those teams to veto or alter projects that pose undue risk. For example, if a social media firm values the mental health of its community (an ethical stance), it might discontinue an algorithm that maximizes engagement but also amplifies anger or envy among users. Instead, it would invest in alternative approaches that promote healthier interactions. This reflects placing user well-being above short-term gains.

- **Inclusivity and Diversity:** Another key organizational value is inclusivity, both in the workforce and product design. Diverse teams are better positioned to anticipate a wider range of risks and create more universally accessible solutions. If a company values diversity, it will ensure that voices from different backgrounds (including those representing vulnerable groups) have a seat at the table when AI products are being developed. This guards against blind spots that could lead to biased or unsafe outcomes. For government organizations or public institutions, valuing inclusivity means actively engaging diverse communities in policymaking around AI.
- **Service and Purpose:** Particularly for public institutions (but also for mission-driven companies), valuing service to society is important. For example, a healthcare organization deploying AI diagnostic tools should prioritize patient dignity and informed consent over efficiency alone. A government deploying AI in public services should value equity and citizen rights above cost savings. When organizations frame themselves as servants of a greater purpose (be it improving education, connecting people, or empowering creativity), they keep themselves aligned with societal values. Klaus Stoll and Sam Lanfranco suggest making human rights a part of the business plan (Stoll and Lanfranco 2023)—that is, essentially urging organizations to adopt the value of human well-being as a corporate priority, not just a CSR sidebar.

In practice, many organizations articulate values like “integrity,” “customer focus,” or “innovation.” The challenge is ensuring that human-centric values, such as integrity, transparency, and respect for users, are embedded in incentives and decision-making processes. One actionable step is tying executive compensation or team evaluations to ethical outcomes (for instance, a bonus metric for reducing incidents of AI harm or increasing measures of customer trust). This cements the idea that these values are not just lip service but real operational priorities.

5.2 Societal Values (National or Cultural Values)

At the level of society, values are the collective principles that shape communities and nations. These values are reflected in laws, cultural norms, and public rhetoric. For example, many democratic societies value freedom, equality, and human rights, whereas other societies might prioritize values like harmony, community, or tradition. In the context of AI, societal values influence what regulations are passed and what behaviors are considered acceptable or abhorrent. We've touched on some societal values in previous sections, like dignity and justice, which certainly apply here. Let's highlight a few:

- **Human Dignity:** As a societal value, dignity means that every person is treated with respect and their intrinsic worth is recognized. Timothy Shriver emphasizes that dignity and respect "are what we all deserve," recalling how mid-twentieth-century America experienced a "fundamental…awakening of dignity" that fueled social reform (Shriver 2025). Today, debates about AI often hinge on dignity. For instance, is it acceptable for AI systems to monitor workers continuously for productivity? Or does that treat humans like machines? A society that values dignity may restrict such practices, insisting that technology must not degrade human experience. The Dignity Index Shriver mentions is an attempt to score public language on dignity versus contempt (Shriver 2025). Similarly, we can imagine scoring our technologies or policies on whether they uphold or undermine human dignity.
- **Justice and Rights:** Societies (especially liberal democracies) put a high value on justice, rule of law, and individual rights. These values directly shape AI governance. For example, because we value privacy, we institute privacy laws (as seen in the EU's GDPR or various state laws). Because we value free and fair elections, there's increasing insistence that AI political ads be labeled or deepfakes be banned in election context. Because we value equality, society pushes for diversity in tech and fairness in algorithms. If a society's values are strongly against discrimination, you will see strong reactions (public outcry, activism, litigation) when an AI system is found to be racist or sexist. The cultural conversation in the US and elsewhere about biased AI in policing or hiring comes from this value source. Thus, societal values act as a North Star that policy should align with. Lawmaker Aisha Wahab's drive to protect women and children from AI exploitation (Wahab 2024) reflects Californian societal values around protecting the vulnerable and ensuring justice for victims.

- **Progress and Innovation (Tempered by Responsibility):** Many societies, especially in the modern era, value scientific and economic progress. This can sometimes conflict with caution. However, a nuanced societal value that's emerging is responsible innovation. People do want the benefits of AI: better healthcare, convenience, and economic growth, but public opinion often stresses it should be "done right." Societies value safety, as we discussed, so there's broad support for guardrails. A telling example: surveys of the public (including younger generations) show they are wary of certain AI developments. Taylor Berry noted that a huge majority of Gen Z is actually quite critical of the internet's impact on them (Berry 2025). Eighty-five percent think it's too much and harms them. This indicates a societal undercurrent valuing moderation and balance over unchecked tech immersion. It suggests society (at least that generation) values a progress that does not "ruin their soul," a more human-centered progress. Policymakers respond to these values. For instance, we see proposals for rights such as the "right to disconnect" from work in some countries, acknowledging that 24-7 tech connectivity shouldn't override rest, a value important for health and family.
- **Solidarity and Community:** Another societal value, more pronounced in some cultures, is solidarity, caring for each other and not leaving anyone behind. This value can shape how we approach AI's impact on jobs or inequality. If solidarity is valued, a society might invest heavily in retraining programs, or it might slow automation in certain sectors until workers can transition. Or, if community is valued, there might be public support for regulating AI social networks that are seen to be eroding community bonds or sowing discord. On the flip side, societies that value individualism heavily might approach these issues differently, perhaps putting more onus on personal responsibility than collective action. But even individualistic societies like the US have lines they draw for community protection (e.g., against child exploitation, as we've seen).

It's worth noting that societal values are not monolithic or static. They evolve and are often contested. Part of the current discourse around AI is essentially a values negotiation: How much do we value privacy versus security? Free expression versus protection from harm? Innovation versus precaution? These debates play out in legislatures and media. The outcomes will define each society's approach to AI.

One striking societal value invocation is Guterres's appeal to harness AI to "bridge divides" and accelerate development for all (United Nations 2023). Implicit is the value of global solidarity. We should use AI to reduce global inequities (like lack of healthcare or education in poorer regions) rather than widen them. So societal values can also be global values, as the world becomes more interconnected.

5.3 Individual Values

Finally, at the individual level, we consider the values each person holds dear and how those influence their interaction with AI. Individuals vary widely, but generational trends or cultural upbringing shape common patterns. Several individual values are particularly relevant:

- **Personal Well-Being and Mental Health:** Many individuals, especially after experiencing the first wave of digital overload, are reasserting the value of their own well-being. Taylor Berry's article highlights that Gen Z individuals prioritize being happy and mentally healthy, listing happiness (65 percent) and good mental health (49 percent) among their top goals (Berry 2025). This personal value is leading some to change how they use technology; for instance, digital detoxes, seeking more in-person connections. An individual who values mindfulness and peace might choose to limit AI usage (like turning off algorithmic recommendations that lead to doom-scrolling). The market is responding too: there's a rise in apps and tools for focus, for digital wellness, etc., driven by individual demand. When many individuals share a value (e.g., privacy or authenticity), they collectively push products to adapt. For example, the rise of encrypted messaging came from individuals valuing privacy in their personal communications.

- **Autonomy and Agency:** Individuals generally value having control over their lives and choices. If people feel AI systems are making decisions for them in an opaque way (what to watch, what route to drive, even whom to date via algorithms), some pushback occurs because personal autonomy is prized. Kenneth Lim and colleagues discuss visions where learners either become passive "AI operators" or active "AI creators" (Lim, Hilmy, and Wei 2025). Many individuals, especially the young when educated about it, would likely prefer to be active agents with AI, aligning with a value of autonomy and mastery. If I use an AI assistant, I still want to feel in control of the outcomes; I don't want to be a servant to the GPS or the recommender system. This personal value could translate into user behaviors like customizing tech settings or advocating for manual options alongside AI automation. For instance, some drivers value the freedom to drive manually and dislike cars that override them too much, reflecting a value of control.
- **Privacy and Personal Space:** Not everyone values privacy equally, but many do treasure a certain level of personal space and data protection. You'll find individuals who cover their laptop cameras or reject smart home devices because they value their privacy. This personal value can lead them to make different consumer choices, and collectively, that can influence the market (like the emergence of privacy-centric tech products). Also, individuals value the sanctity of personal relationships. Note how Gen Z said in-person relationships are more valuable than digital (Berry 2025). So an individual might choose to put away the phone during dinner to honor that personal value of presence. When multiplied, such individual choices can reshape how tech is used (e.g., norms around phone use in social settings).

- **Integrity and Authenticity:** Many individuals hold themselves to moral principles (don't lie, be kind, treat others as you wish to be treated). These values carry into online behavior too. A person who values honesty will be troubled by deepfakes and may refuse to share unverified information. Someone who values kindness will refrain from trolling and may call out online abuse when they see it. The sum of individual ethical choices contributes to the digital culture. Shriver's call for each person to "treat people with dignity" (Shriver 2025) is a direct appeal to individual values. If we all did this in our daily interactions, online and offline, the internet would be a far less toxic place. He gives examples of individual actions (his mother teaching swimming to disabled kids, Peace Corps volunteers serving abroad) driven by compassion and dignity values, which collectively made society better (Shriver 2025). In the AI context, an individual's values might influence how they choose to use AI; for example, an artist who values authenticity might opt not to use AI generators to replace their creative process, or an employee who values fairness might speak up if they see their company's AI doing something biased.

When individual, organizational, and societal values align, that's when true progress guided by the Human North Star happens. For example, if individuals value privacy, societies enforce it through regulation, and organizations design for it, then AI products will reflect that alignment. Misalignment, however, causes friction. Consider a case where an individual worker wants to embed ethics, but their organization only values profit, and the societal regulations are weak. That worker faces a moral dilemma, and likely, the product will lack ethical considerations.

To improve alignment:

- Organizations can engage with society (public consultations, integrating societal values in their charters).
- Individuals can be educated about the societal impact of their tech usage (so they internalize bigger-picture values too, not just convenience).
- Societies can hold organizations accountable and empower individuals (through rights and education).

One interesting thread from The Register article is how Haugen suggests adopting safety practices from other sectors: car companies don't cut corners on safety because they know independent tests will expose them (Claburn 2024). If society demands such tests for AI (societal value: accountability), and organizations accept and prepare for them (organizational value: responsibility), then individuals benefit and trust the systems more (individual value: trust, security).

In summary, the values we hold dear at each level are interdependent.

- Organizations must broaden their priorities beyond profit to include ethics, transparency, and user-centric ideals. A company that declares its mission to improve the world with AI should, as part of its core values, pledge to do so responsibly and inclusively. Government bodies likewise should value citizen trust and participation, not just efficiency.
- Society must clarify the values it expects technology to uphold, be it freedom, dignity, equality, or others, and encode those into norms and laws. Currently, there's a strong societal stance forming around values like human rights in AI (Stoll and Lanfranco 2023), which is promising. Social movements and public opinion will continue to shape how hard or soft those values feature in policy.
- Individuals ultimately anchor it all; their daily choices and voices set the demand. As more people like Generation Z explicitly call out the soul-sapping aspects of the current internet and long for "quiet, clarity, depth and connection," things "that can't be found in an algorithm" (Berry 2025), they are signaling a personal reorientation toward more profound values. This will inevitably push tech to adjust, as user preferences always do.

In navigating AI's future, it might help to occasionally pause and ask at each level: Are we being true to our best values? For a person, am I treating others online with respect? For a company, are we putting people before profit where it really counts? For a society, are our laws reflecting what we deeply cherish about humanity? These reflections ensure that our Human North Star doesn't get lost amid the excitement or fear surrounding AI. If we hold fast to our cherished values at all levels, they will act like a compass needle, always pointing toward the ethically correct direction, even when external conditions (market pressures, geopolitical tensions, etc.) threaten to knock us off course.

6.

How Do We Promote the Flourishing of Humanity, and What Will It Look Like?

Ultimately, the goal of navigating the AI revolution is not just to avoid hazards but to arrive at a better place: a future where humanity flourishes with the help of AI. Human flourishing is a comprehensive concept. It implies people living in a state of overall well-being, fulfillment, and opportunity, both individually and collectively. It is a future where needs are met, potentials are realized, and new horizons of achievement and understanding are reached. The question then is twofold: How do we actively promote such flourishing using AI, and what might a flourishing society augmented by AI actually look like in concrete terms?

Reaching this future is not automatic. It requires sustained investment, political coordination, and strong governance, and it will face real friction, such as unequal access to data and compute, misaligned incentives, and uneven implementation capacity. Without deliberate policy and funding choices, benefits can concentrate among a few while harms fall on the vulnerable. These constraints make it even more important to pair innovation with accountability, transparency, and equitable deployment.

First, consider the promotion of flourishing. This involves deliberate efforts, policies, and designs to ensure AI is leveraged for positive ends. It echoes Guterres's rallying cry for a "race to develop AI for good," flipping the narrative from racing for supremacy or profit to racing for human well-being (Guterres 2023). There are several key avenues to promote flourishing.

6.1 Direct AI Toward Solving Humanity's Greatest Problems

To foster flourishing, we should aim the powerful tool of AI at the issues that, when solved, most improve human lives. The United Nations' Sustainable Development Goals (SDGs) provide a roadmap of such problems: poverty, hunger, disease, lack of education, inequality, climate change, and more.

Antonio Guterres highlighted that AI, if used well, can "end poverty, banish hunger, cure cancer, supercharge climate action," and accelerate achievement of the SDGs (Guterres 2023). Promoting flourishing means prioritizing these uses.

Governments and private entities can fund AI research in these domains and create incentives such as prizes or subsidies for breakthroughs. For example, an AI system that greatly improves crop yields for small farmers, or AI that can predict and prevent disease outbreaks.

We are already seeing glimpses of this future: AI systems that analyze medical scans faster and more accurately, potentially saving lives through early detection of illnesses; climate models powered by AI that improve disaster preparedness; AI tutors that reach children without access to quality schooling, thereby inching toward education for all.

In a flourishing scenario, these innovations would not remain isolated pilots but would become widespread and accessible. A world where AI has helped eliminate certain diseases, where no one goes hungry because agriculture is optimized, and where learning is tailored to every child's needs, that is a picture of human flourishing advanced by technology.

6.2 Ensure AI Augments Rather Than Replaces Human Capacities

Flourishing is about humans thriving, not machines thriving at humans' expense. Kenneth Lim and colleagues present two contrasting visions in education: one where students are passive recipients of AI outputs, and another where students become creative agents harnessing AI (Lim, Hilmy, and Wei 2025). The second vision is clearly more aligned with human flourishing. It nurtures "active creators who harness technology for community benefit" (Lim, Hilmy, and Wei 2025).

Expanding this beyond education, in a flourishing future, AI serves as a tool that empowers people to achieve more, rather than making people obsolete or dependent. For instance, consider work: rather than AI causing mass unemployment, a flourishing-centric strategy would redesign jobs and economic structures so that AI takes over mundane tasks and humans move into more creative, strategic, or caregiving roles—roles that draw on empathy, creativity, and complex problem-solving.

This may require bold policy interventions, such as large-scale reskilling programs or a shorter work week supported by AI-driven productivity gains, giving people more time for family, community, and creative pursuits.

A flourishing humanity is one in which work is more fulfilling on average because drudgery is minimized and human contribution is valued in distinctly human terms, often in collaboration with AI.

6.3 Foster a Digital Environment That Enriches Human Experience

In a flourishing scenario, our digital lives contribute positively to overall happiness and social cohesion rather than detracting from them. Social media and communication platforms powered by AI could be reoriented to strengthen relationships and understanding. Timothy Shriver asks poignantly, “Will AI make us happier and more likely to flourish?” and finds that leaders often say they have “no idea” (Shriver 2025). We must turn that uncertainty into intentional design for joy and dignity.

Imagine social AI systems that, instead of exploiting anger and division, actively promote uplifting content, cross-cultural dialogue, and community-building. This might sound idealistic, but these platforms are human-made. They can be tweaked to prioritize “heart,” as Shriver’s friend Angelo Moratti suggests (Shriver 2025).

A concrete example: AI could help connect lonely individuals with others of similar interests in safe, moderated environments, effectively helping to address the loneliness epidemic by creating genuine friendship opportunities. VR and AI together might enable rich telepresence experiences, letting families spread across the globe feel together.

Flourishing, in this sense, means that technology enhances emotional and social well-being rather than fragmenting it. Shriver’s emphasis on dignity is critical here—a flourishing digital environment is one where people treat each other with dignity online, and where AI systems actively filter out dehumanizing content and reduce toxic interactions (Shriver 2025).

6.4 Cultivate Enlightenment and Creativity

Human flourishing extends beyond material comfort to include intellectual, cultural, and even spiritual growth. AI can be a tremendous enabler here. Think of an AI that can compose music alongside a human artist, sparking new creativity, or an AI that can simulate historical worlds for learners to explore, vastly enriching education and empathy for other cultures.

John Kennedy Philip describes transhumanist optimism about overcoming human limitations like aging or limited intelligence (Philip 2024). While some of those visions are controversial, they point to a desire to expand human potential. Short of radical transhumanism, we can imagine AI helping humans access knowledge instantly, learn new skills with personalized training, or even enhance cognitive abilities. For instance, an AI assistant that helps someone with memory difficulties or provides real-time language translation in one's ear, breaking barriers to communication. If done ethically, these enhancements mean individuals can flourish more fully, spending more time in the pursuits that give them meaning and less on trivial obstacles.

However, as Philip (2024) noted through Fukuyama's concerns, we must be careful that such enhancements don't cross into violating human rights or creating a new elite vs. underclass (a non-flourishing scenario). In a flourishing vision, access to beneficial enhancements or AI aids would be equitable and would respect the essence of being human (e.g., not taking away free will or authentic emotion).

6.5 Embed the Value of Dignity and Meaning in All AI Applications

Flourishing is closely tied to dignity and a sense of meaning in life. Timothy Shriver imagines that treating people with dignity can be "a history-changing force again" for our times (Shriver 2025). So how do we ensure AI promotes dignity? Partly by what we choose to do with AI (as covered above in solving big problems and augmenting people) and partly by what we choose not to do.

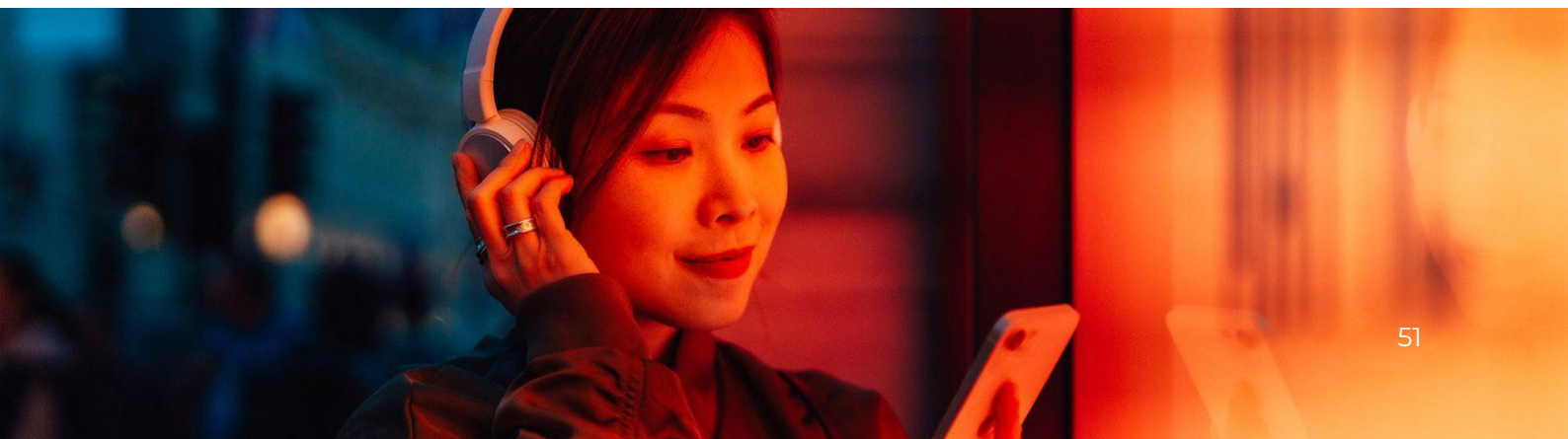
A flourishing-oriented roadmap likely means rejecting certain uses of AI that are antithetical to human values, even if they offer some efficiency gains. For example, a society aiming for true human flourishing might decide not to implement AI systems that impose inhumane control (like scoring citizens or pervasive surveillance that chills freedom) because those strike at dignity and agency. Instead, it opts for uses that liberate and uplift.

We've seen how California decided that some uses (deepfake porn) are simply beyond the pale (Wahab 2024), as they should be in any human-centric future. Flourishing also means individuals have purpose. AI could assist in that by taking over survival-level concerns so humans can engage in higher pursuits: arts, exploration, volunteering, and learning. One could imagine a scenario where, with AI-driven productivity, society shifts to guarantee basic needs and people spend more time in lifelong learning and community activities, leading to a more enlightened populace.

Drawing together the values, rights, guardrails, and frameworks explored throughout this paper, we can now envision what a flourishing AI-assisted future might look like in daily life.

- **Morning:** You wake up well-rested because your AI-managed home subtly adjusted the temperature and lighting for optimal sleep. You check your health stats via an AI health guardian that overnight monitors your vitals (with full privacy protection) and perhaps detects an issue early and schedules a checkup. Preventive healthcare is rampant, so diseases are caught early or even predicted and prevented thanks to AI analysis of your personal health data in a secure way. The result: people live longer, healthier lives (the promise of life-extension that transhumanists talk about, but achieved in a humane, equitable way rather than only for the rich).
- **Work/Education:** If you're a student, you log into a virtual classroom with an AI tutor that knows your strengths and weaknesses intimately (Lim, Hilmy, and Wei 2025). It guides you through a customized curriculum at your pace but also encourages you to create projects and collaborate with other students worldwide, cultivating creativity and teamwork. Education is no longer one-size-fits-all; it's personal, engaging, and teaches not just facts but critical thinking, with AI prompting you to reflect and not just consume answers (Lim, Hilmy, and Wei 2025). If you're a worker, perhaps you've seen routine tasks largely automated, but rather than unemployment, this has given you more time to focus on strategy, interpersonal aspects, or innovation in your job. Maybe you work fewer hours for the same pay because society reaped productivity gains. Or you've transitioned careers with help from an AI career coach that identified your transferable skills and arranged training in a field of your interest. People don't fear AI taking jobs because they see a robust support system and many new opportunities arising, ones that play to human strengths like imagination, empathy, and complex problem-solving.

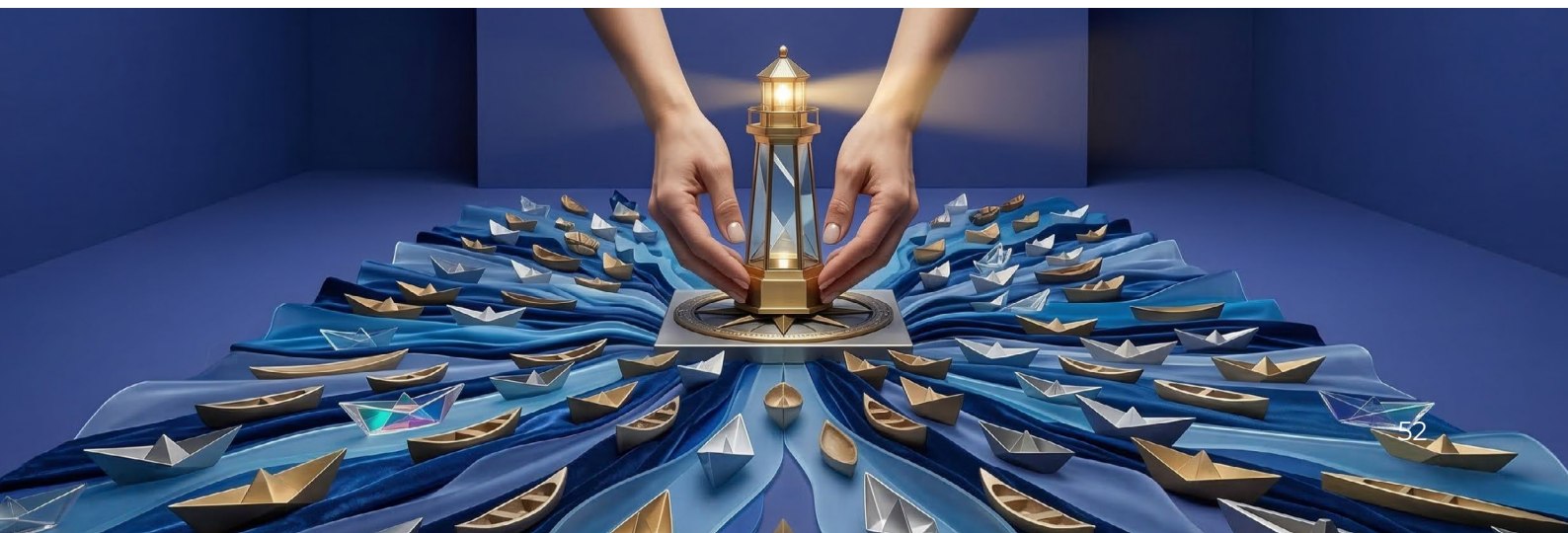
- **Community and Civic Life:** Throughout the day, AI background services quietly enhance civic life, and traffic flows smoothly due to AI coordination, reducing daily stress and pollution. Public safety is improved, but in a balanced way. For example, AI surveillance might exist but is governed transparently with community oversight, and crime prevention AI programs address root causes, such as identifying areas for social services, rather than simply punishing. If you interact with the government, say to get a permit or benefit, AI chatbots handle routine queries instantly but always offer a human handoff if needed, and they are designed to be clear and respectful (the government values citizen experience, not making bureaucracy a nightmare). Democratic participation is bolstered. You might get AI-curated briefings on local issues tailored to be informative and unbiased, helping you make better decisions in town hall meetings or polls. Misinformation is much less of a scourge because AI filters, combined with digital literacy, have largely contained it, and media AI is more likely to connect you with constructive news or stories from diverse perspectives, increasing understanding across different groups.
- **Evening and Leisure:** With more leisure time and a society that values balance, you pursue hobbies enhanced by AI. Perhaps you paint with an AI that offers new techniques to try, or you play music with AI-generated accompaniment, making creative expression accessible to even those without formal training. Or you simply spend time with loved ones, and AI takes care of domestic chores via robotics, freeing you to bond and relax. Entertainment is immersive and personalized, but also communal. Maybe virtual reality gatherings are common, letting families across continents have dinner “together” in a simulated space that feels real. Importantly, you have the choice to disconnect. Technology doesn’t coerce your attention. Many people choose to be offline at times to enjoy nature, as the culture has recognized the value of that balance (Berry 2025). AI helps environmental flourishing too. Cities are greener due to smart energy and climate adaptation projects guided by AI; climate change has been mitigated to an extent by AI-optimized systems finding efficiencies and new solutions for carbon capture.



- **Lifelong and Societal Outcomes:** Stepping back, society in this vision has addressed many of the chronic issues. Poverty is drastically reduced because AI-driven productivity and better governance have increased abundance and improved distribution. Perhaps something like a universal basic income or services exists, funded by the wealth AI creates, so no one lacks basics. People use that security to educate themselves or start enterprises, further adding to a virtuous cycle. Global inequality is smaller. Developing countries benefitted from AI tech transfers and no longer suffer extreme shortages, partly due to global coordination—a result of the international governance Guterres advocated (Guterres 2023). New challenges, of course, will arise (they always do), but humanity has more tools and more wisdom to tackle them.

This scenario may sound idealized, but it serves as a beacon—an image of human flourishing with AI. It's not utopia. There will still be debate, diversity of lifestyles, and the eternal challenges of human relationships and meaning. But overall, it's a world where AI is a partner in human progress, not a threat.

To achieve this, we need conscious effort today. Timothy Shriver's question "How will you bring heart to technology?" is apt (Shriver 2025). Heart implies empathy, compassion, human-centric thinking, essentially injecting our North Star values into every step of AI development and deployment. When tech CEOs, engineers, policymakers, educators, and all of us keep that question in mind, we start to design for flourishing. For instance, an engineer might decide not only to measure an algorithm's accuracy but also its impact on user happiness or community resilience. A policymaker might not just think about economic growth from tech but also whether it increases the net dignity and solidarity in society. A teacher might embrace AI in the classroom only in ways that encourage students' growth, not passivity.



There is also a philosophical backdrop: John Kennedy Philip observed that transhumanists see conquering aging and expanding intellect as key to maximizing benefit to humanity (Philip 2024). Whether or not we go to those extremes, the impulse is to break former limitations. In a flourishing future, many things once thought impossible for humanity might become reality: cures for diseases like cancer could indeed be found (a major source of human suffering gone), average lifespans could extend with quality (raising new societal questions but also joys), and we might solve currently intractable problems like clean energy through AI-driven innovation. It's important, though, as Philip's mention of critics reminds, to do this without losing our humanity (Philip 2024). The flourishing we aim for is a human flourishing. AI is a means, not the end. We wouldn't measure success by, say, how intelligent or autonomous AI itself becomes (like a sci-fi scenario where AI surpasses us). We measure it by human outcomes: health, happiness, knowledge, virtue, and fulfillment.

In conclusion, promoting the flourishing of humanity with AI is about intentional guidance and ethical use of technology to enhance human life in every dimension. What it will look like is a world of solved problems and open opportunities, a world where AI is so seamlessly integrated and benevolent that it fades into the background, and what we notice is not the AI itself but the greater freedom, creativity, and solidarity we experience. It's a world where technology, guided by the North Star of human values, truly serves "the common progress of mankind" (to quote a value from the UN ethos). Achieving this vision isn't guaranteed, but it is within our grasp if we commit, collectively, to the principles discussed throughout this paper: shared values, rights, collaboration, guardrails, and alignment at all levels of society. The open seas of AI need not be feared. With our Human North Star, we can navigate to a future where all of humanity can flourish on these new horizons.

Conclusion

The dawn of the AI age finds humanity at a crossroads, charting a course into unfamiliar waters. In this white paper, we have identified a constellation of guiding questions, values, rights, progress, safety, levels of values, and human flourishing that together serve as the North Star for our journey. Through careful analysis of expert insights and real-world developments, several overarching themes emerge:

- **Human-Centric Guidance:** AI must be developed and deployed with steadfast reference to human values and rights. Our shared values, dignity, compassion, fairness, and accountability, are not abstract ideals but practical compass points. They inform clear inalienable rights that AI systems should respect, from privacy and non-discrimination to the right of individuals to maintain agency and dignity in a digital world. Keeping these values and rights front and center is the surest way to avoid drifting off course. As Frances Haugen cautioned, without a guiding North Star, even well-intentioned organizations can lose their way and erode trust by prioritizing short-term gains (Haugen 2024). By contrast, a human-centric approach enshrines "people over profit" as more than a slogan; it becomes the operational ethos of innovation.
- **Collective Stewardship and Collaboration:** Navigating the "open seas" of AI is a collective endeavor. International cooperation, multi-sector partnerships, and inclusive dialogue are indispensable for setting global guardrails and norms. Whether through a UN-led body for AI governance (Guterres 2023) or industry coalitions committing to ethical standards, working together amplifies our ability to steer effectively. No single nation or company can address the far-reaching impacts of AI alone. Our analysis underscored that collaborative efforts, from sharing best practices on AI safety to jointly funding AI-for-good projects, will accelerate progress toward the future we actually want. In essence, collaboration is our compass correction mechanism. It helps humanity stay unified in direction despite different interests and perspectives. It transforms AI from a zero-sum competition into a shared voyage with all hands on deck.

- **Proactive Guardrails for Safety and Equity:** The journey toward AI-enhanced growth requires sturdy guardrails to prevent crashes and to protect those at risk. We detailed how legislative, technical, and ethical guardrails can mitigate threats, be it laws banning malicious deepfakes (Wahab 2024), trustmark frameworks to certify human-rights-respecting AI (Stoll and Lanfranco 2023), or organizational transparency that invites public scrutiny (Claburn 2024). A recurring insight is that safety and ethical considerations must be built in from the start, not retrofitted after disasters. It is far easier to design AI platforms with content moderation, privacy protection, and bias mitigation upfront than to fix wreckage after harm has occurred. Guardrails are not the enemy of innovation; they are enablers of sustainable innovation. They ensure that AI's benefits do not come at the price of social instability or moral compromise. By especially safeguarding vulnerable populations, children, minorities, and those lacking power, we uphold the principle that technological progress should lift up all of humanity, not just the privileged or powerful.
- **Value Alignment Across Levels:** A key to consistency and integrity in our approach is aligning values at individual, organizational, and societal levels. When individuals practice digital civility and demand ethical tech, organizations feel pressure to respond, and society gains momentum for normative change. When companies adopt responsible innovation values and governments legislate in line with public conscience, individuals are empowered to flourish without having to swim upstream against toxic digital currents. This alignment—the harmonization of personal ethics, corporate responsibility, and cultural norms—creates a reinforcing loop, a virtuous cycle steering us toward our North Star. Conversely, misalignments (like a society that values privacy but a company that does not, or vice versa) create friction that must be resolved through dialogue and, if needed, regulation. Our exploration shows positive signs. For instance, public concern over tech's effects on mental health (Berry 2025) is pushing platforms to reconsider addictive design while insiders like Haugen have amplified societal calls for transparency (Haugen 2024). These are alignment processes in action.

- **Intentional Vision of Flourishing:** Finally, we must not lose sight of the destination: the flourishing of humanity. AI is not an end in itself; human well-being is. By envisioning in concrete terms what a better future looks like—healthier lives, liberated creativity, stronger communities, expanded knowledge, and a preserved planet—we give ourselves a target to strive for. Antonio Guterres's vivid appeal for AI to "cure cancer…and propel us toward the SDGs" (Guterres 2023) and Timothy Shriver's plea to infuse technology with heart and dignity (Shriver 2025) both anchor our aspirations in tangible human outcomes. A future where AI helps end hunger or educates every child is a future where technology has become a true servant of the human family. It is a future where the long arc of innovation bends toward justice and joy. Achieving it requires resolve, resources, and moral courage, but our analysis affirms it is within reach, provided we make the choices now that keep us on the path.

In navigating uncharted waters, sailors of old would continually fix their gaze on the North Star to correct any drift. So too must we, as we sail the seas of AI. There will be storms, controversies, setbacks, and unintended consequences, but our guiding light remains the constellation of human values and rights we collectively affirm. We have tools that our predecessors lacked: the ability to connect globally in real time, unprecedented computational power to model scenarios, and a growing repository of lessons from the early internet and social media era. With these, we can anticipate challenges and chart wiser courses.

It is also important to acknowledge humility on this journey. We do not know everything about AI's potential or perils. That is why adaptability, continuous learning, and inclusiveness of diverse voices are so vital. If one thing is certain, it is that the journey will surprise us. But armed with a North Star, surprises need not throw us off course; they can become merely new data points for refinement rather than existential threats.

To conclude, Navigating the AI Open Seas: The Human North Star is more than a concept. It is a call to action and a framework for accountability. It reminds every stakeholder, the coder, the CEO, the regulator, the educator, the everyday user, that we each hold a piece of the compass. By adhering to our shared values, upholding inalienable rights, collaborating for the common good, enforcing prudent guardrails, aligning our actions with our principles, and ever focusing on human flourishing as the true measure of success, we can ensure that AI serves humanity and not the other way around. The open seas of AI are vast and deep, but with a steadfast North Star, we can navigate them with confidence.

In the words of Frances Haugen, “we can work together…and shape a future that is going to work for all of us, where we’re not afraid to innovate…and make the most impact out of these technologies” (Claburn 2024). That future—one of promise, anchored in our highest ideals—is one we all have a hand in creating. The course is set. Let us sail forth guided by the enduring light of our human North Star.

References

1. Amodei, D. 2025. “Machines of Loving Grace.” DarioAmodei.com. March 15. https://www.darioamodei.com/essay/machines-of-loving-grace.
2. Berry, Taylor. 2025. “85% of Gen Z Thinks the Internet Is Ruining Their Soul—and They’re Right.” Relevant Magazine, May 30. https://relevantmagazine.com/current/oped19/85-of-gen-z-thinks-the-internet-is-ruining-their-soul-and-theyre-right/.
3. Claburn, Thomas. 2024. “Facebook Whistleblower Calls for Transparency in Social Media, AI, Haugen Says Navigating the Digital World Requires a North Star.” The Register, August 27. https://www.theregister.com/2024/08/27/facebook_transparency_ai/.
4. diVittorio, Alicia. 2024. “Frances Haugen’s 2024 DataGrail Summit Keynote: Lessons from the Past to Drive Responsible Innovation.” DataGrail Blog, September 9. https://www.datagrail.io/blog/data-privacy/frances-haugens-2024-datagrail-summit-keynote-lessons-from-the-past-to-drive-responsible-innovation/.
5. Guterres, António. 2023. “Remarks at the UN Security Council Debate on Artificial Intelligence.” United Nations, July 18. https://press.un.org/en/2023/sgsm21880.doc.htm.
6. Haugen, Frances. 2024. Keynote Address at DataGrail Summit: “Responsible & Respectful Innovation,” Half Moon Bay, CA, August 27. https://www.youtube.com/watch?v=C8SqwhwMPvI.
7. Lim, Kenneth Y. T., Ahmed Hazyl Hilmy, and Bryan Kuok Zi Wei. 2025. “AI Operators or Creators? Two visions of agency and learning.” UNESCO Future of Education Ideas, May 26. https://www.unesco.org/en/articles/ai-operators-or-creators-two-visions-agency-and-learning.
8. Philip, John Kennedy. 2024. “A Philosophical History of Transhumanism.” Philosophy Now, Issue 160 (February/March) https://philosophynow.org/issues/160/A_Philosophical_History_of_Transhumanism.

9. Shriver, Timothy. 2025. “We Must Do Better.” The Dignity Index via Maria Shriver’s Sunday Paper, June 14. https://www.mariashriversundaypaper.com/tim-shriver-treating-people-with-dignity/.
10. Stoll, Klaus, and Sam Lanfranco. 2023. “Realizing the Promise of AI in a World Where Human Rights Matter.” CircleID (Digital Citizen), October 13. https://circleid.com/posts/20231013-realizing-the-promise-of-ai-in-a-world-where-human-rights-matter.
11. Wahab, Aisha. 2024. “State Senator Dr. Aisha Wahab’s 2024 Full AI Safety Package Signed into Law.” Press Release, California State Senate (District 10), September 29, 2024. https://sd10.senate.ca.gov/news/state-senator-dr-aisha-wahabs-2024-full-ai-safety-package-signed-law.

Authors

Mickie Chandra

Mickie serves in various capacities with advocacy organizations at the US local and state levels and is a Senior Executive Fellow at The Digital Economist. Her experience includes launching and driving successful grassroots initiatives that support the K–12 education sector, focusing on technology and economic development, leading programs for K–12 students in the arts and STEM fields, building systems to engage communities on pressing issues, and applying whole-systems thinking to develop viable solutions and competitive advantage. She has also built successful cross-disciplinary partnerships and served on state task forces focused on technology and education. Mickie is deeply committed to societal impact through the development of human-centered systems and policies, and she promotes the principle that community growth and economic vitality should remain human-centric.

Nikhil Kassetty

Nikhil Kassetty is a technology leader, speaker, and author focused on AI-enabled systems, financial technology, digital trust, and scalable software architecture, and is an Executive Fellow at The Digital Economist. With over a decade of experience building enterprise platforms and modernizing complex technology ecosystems, he operates at the intersection of innovation, governance, and real-world impact. His work spans AI, fintech, cloud-native engineering, and responsible technology adoption, with a particular focus on aligning emerging systems with human values, accountability, and long-term societal benefit. In addition to his technical leadership, Nikhil contributes to broader conversations on ethical AI, human-centered innovation, and the future of digital systems through writing, speaking, mentoring, and cross-sector collaboration. His work bridges practical implementation and strategic vision, helping shape technology that is both scalable and socially responsible.

Contributors

Nithin Singh Mohan

Nithin Singh Mohan is an AI and supercomputing leader at Hewlett Packard Enterprise, where he builds and scales advanced AI systems at supercomputing scale. He brings extensive experience spanning enterprise AI, high-performance computing, and startup innovation, including leadership roles within unicorn-stage companies. As a Senior Executive Fellow at The Digital Economist, Nithin contributes thought leadership on the future of AI infrastructure, agentic systems, and compute-driven innovation. His work bridges deep technical expertise with strategic insight, helping organizations translate large-scale AI capabilities into measurable business and economic impact.

Melissa Tony Stires

Melissa Tony Stires is a global speaker, technology leader, and advocate for human-centered innovation working at the intersection of artificial intelligence, wellbeing, and inclusive growth. As Chief Executive Officer of the Fundamental Wellbeing Foundation and Founding Partner and Chief Global Growth Officer at Mia AI, she focuses on applying emerging technologies to strengthen human connection, trust, and cross-cultural understanding while advancing sustained wellbeing at scale. Her work bridges consciousness research, evidence-based practices, and AI-driven solutions, and she regularly collaborates with global leaders, policymakers, and industry executives on strategic initiatives and partnerships. A Senior Executive Fellow at The Digital Economist, Melissa contributes to advancing dialogue on human-centered AI and inclusive leadership, bringing a globally informed, values-driven perspective to how technology can support meaningful societal outcomes.

About

The Digital Economist, headquartered in Washington, D.C. with offices at One World Trade Center in New York City, is the world's foremost think tank on innovation advancing a human-centered global economy through technology, policy, and systems change. We are an ecosystem of 40,000+ executives and senior leaders dedicated to creating the future we want to see—where digital technologies serve humanity and life.

We work closely with governments and multi-stakeholder organizations to change the game: how we create and measure value. With a clear focus on high-impact projects, we serve as partners of key global players in co-building the future through scientific research, strategic advisory, and venture build out.

We engage a global network to drive transformation across climate, finance, governance, and global development. Our practice areas include applied AI, sustainability, blockchain and digital assets, policy, governance, and healthcare. Publishing 75+ in-depth research papers annually, we operate at the intersection of emerging technologies, policy, and economic systems—supported by an up-and-coming venture studio focused on applying scientific research to today's most pressing socio-economic challenges.

CONTACT: INFO@THEDIGITALECONOMIST.COM

CENTER OF EXCELLENCE

The Digital Economist Center of Excellence for a Human-Centered Global Economy is dedicated to addressing the biggest challenges humankind and our planet face by leveraging digital technologies for good.

The Digital Economist Executive Fellowship invites senior leaders and decision-makers to join our Center of Excellence, providing them with a platform for amplification and global impact. This unprecedented, one-of-a-kind opportunity enables Executive Fellows to network and build relationships at the highest level, driving transformative change and innovation in the global digital economy.

The Executive Fellowship

The Digital Economist Executive Fellowship is a selective leadership program integrating visionary professionals into the Center of Excellence for a Human-Centered Global Economy to advance global economic policy and systems transformation.

Global Impact

Amplify your influence and drive transformative change by participating in high-level initiatives that address the most pressing global challenges.

Elite Community

Become part of an exclusive network of visionary leaders and innovators, collaborating to shape the future and drive global progress.

Unparalleled Opportunities

Access unique platforms and events that enhance your professional journey, providing unparalleled opportunities for growth, visibility, and leadership.

Participation Framework

Time Commitment

Minimum commitment: 24 hours per year, for the monthly Center of Excellence meetings. On-demand consultation with the Fellowships team.

Publications

Executive Fellows are expected to contribute to two key publications per year, launched at key global events such as Davos and New York Climate Week.

In-Person Convenings

Executive Fellows are invited to in-person convenings in North America and Europe, with regional convenings in Africa, Latin America, and Asia.

Speaking Engagements

Executive Fellows are offered speaking opportunities throughout the year to amplify their work and contributions.

Our Executive Fellows are at the forefront of research, policy discourse, and systems-level transformation.

- Applied Artificial Intelligence
- Digital Assets & Blockchain
- Sustainability in Tech
- Tech Policy & Governance
- Quantum Computing
- Cyber Studio
- Regenerative Digital Infrastructure
- Healthcare Innovation

Publications

Ideas that shape the future.

The Digital Economist's publications translate research into high-signal outputs: frameworks, policy papers, and industry outlooks that advance a sustainable, inclusive digital economy and inform decision-making across markets and institutions.

Explore our full portfolio of publications and research outputs:
www.thedigitaleconomist.com/publications

Engagement Opportunities

Executive Fellows have access to over **500 events globally** in a Fellowship cycle.

2026 World Bank Group / IMF Spring Meetings
April 13–18, 2026, Washington, DC

The Digital Economist Virtual Summit
June 2026

UN General Assembly (UNGA 81)
September 8–22, 2026

New York Climate Week
September 20–27, 2026

The Digital Economist Virtual Summit
November 2027

Davos Week
January 2027, Davos, Switzerland

Join the Fellowship

Advance your leadership within a global platform shaping technology, policy, and economic systems transformation.

Access Full Brochure | Apply Now | Learn More

Institutional Research Network

A Fragmented World Requires New Institutional Leadership

Technology, economics, and governance are shifting faster than traditional institutions can adapt. AI ecosystems, digital assets, geopolitical competition, sustainability transitions, and new governance architectures demand clarity, legitimacy, and a coherent strategy.

Institutions must now operate as signal generators—shaping the narratives, norms, and systems that define global markets.

Why We Built the Institutional Research Network

A global research and convening platform enabling institutions to:

- Shape emerging policy and governance discourse
- Build narrative power in a volatile environment
- Co-author high-signal research with global experts
- Gain visibility at the world's most influential convenings
- Anchor strategy in human-centered, future-forward frameworks

Co-Authorship & Knowledge Pathways

Through structured co-authorship across eight priority domains—Tech Policy and Governance, Digital Assets & Blockchain, Sustainability in Tech, Applied Artificial Intelligence, Cyber Studio, Quantum Computing, Regenerative Digital Infrastructure, and Healthcare Innovation—institutions contribute to high-level research that informs policy dialogue, regulatory development, and strategic decision-making.

Participation extends beyond commentary. Institutions are integrated into published research, roundtable dialogues, and domain-specific working groups that inform regulatory discussions and industry standards. This structured engagement enables organizations to contribute at the research and drafting stage, engage directly with policymakers and industry leaders, and align internal strategy with emerging policy and market developments, resulting in active presence within decision-making environments rather than passive visibility.

We invite your organization to schedule a strategic briefing to map research priorities and determine the appropriate integration pathway within the Institutional Research Network.

Reach us at partnerships@thedigitaleconomist.com.
Visit us at **thedigitaleconomist.com**

The Digital Economist Ventures

Applied Platforms. Strategic Domains. Real-World Implementation.

Research defines the questions. Ventures test the answers.

In addition to research and convening, The Digital Economist advances a portfolio of venture platforms that extend inquiry into applied domains, where governance, infrastructure, and market design move from dialogue to deployment.

Each venture operates with a defined mandate while remaining integrated within the broader institutional ecosystem.

Tech for Transparency

Financial integrity in the digital age

Advances financial accountability and anti-corruption frameworks through distributed technologies and data-driven transparency systems. Positioned at the intersection of blockchain infrastructure and institutional reform, it translates transparency principles into operational tools.

The Ostrom Project

Reimagining digital commons governance

Explores collective stewardship models for emerging digital systems. Drawing on principles of shared resource governance, it develops frameworks for sustainable digital infrastructure and cooperative system design.

ANER-G

Energy systems innovation

Focuses on decentralized infrastructure, programmable energy markets, and next-generation grid integration. It addresses the structural evolution of energy systems within digital and blockchain-enabled environments.

Africa Coalition

Continental coordination for strategic sectors

Convening leaders across energy, infrastructure, finance, health innovation, education, and future capabilities, the Coalition creates structured engagement pathways for continental collaboration.

Explore the full ecosystem at thedigitaleconomist.com


THE
DIGITAL
ECONOMIST