HRBA@Tech

December 2023

A Human Rights-Based Approach to AI for Tech Startups and Global Normative Governance

A Series of Papers on A Human Rights-Based Approach to New and Emerging Technologies (HRBA@Tech)

2-1 **Applying the HRBA@Tech Model to AI for Tech Startups**

2-2 **Harnessing AI to Solve Climate Change as a "Wicked Social Problem"**

2-3 **The Global Governance Landscape of AI and its Potential to Better Promote a Human Rights-Based Approach to AI**

---

* These papers follow the inaugural report of the series published in 2022: Towards a Human Rights-Based Approach to New and Emerging Technologies: A Framework.

Acknowledgments

This report was co-edited by the Seoul National University AI Policy Initiative (SAPI) and the Universal Rights Group (URG), with Stephan Sonnenberg (SNU), Louis Mason (URG), and Yong LIM (SNU) serving as the editors. Paper 2-1 was co-authored by Stephan Sonnenberg, Yong LIM, Seungbum CHOI, Soo Jin LEE, and Eun Seo JO, all affiliated with SAPI. Paper 2-2 was authored by Stephan Sonnenberg of SAPI. Paper 2-3 was authored by Louis Mason of URG.

We would like to thank Ambassador Seong Deok YUN, who serves as the Permanent Representative of the Permanent Mission of the Republic of Korea in Geneva and Secretaries Youngmin KWON and Woohyun KANG, who commissioned this project and have been unfailingly supportive of this unique transnational collaboration between civil society and academia to help support the development of policies at the global level.

We owe a great debt of gratitude to the Permanent Mission of the Republic of Korea in Geneva for again hosting a roundtable discussion with experts from the United Nations, various diplomatic missions, think tanks, and academic institutions around Geneva. At that event, we had the honor to receive preliminary feedback from numerous diplomatic representatives and policy makers from various UN organizations. We hope to continue such fruitful conversations in the future.

The SAPI team is particularly proud to have been able to involve a group of talented students in this research effort. Daria Chepik, Seunghun CHOI, Hyeon-Woo KIM, Hyunjee KIM, Myoungwon LEE, Seunghyeon LEE, Meike Ly, Kit Pang, Seongho PARK, and Han-Kyoung YU from the Human Dignity Clinic at Seoul National University School of Law provided invaluable research and analytical support for the project, as did Laura Lin and two colleagues from the Yale Lowenstein Human Rights Project. Furthermore, we would like to thank Hajoon KIM, Jeongyeon KIM, and Soo Young YUN, all of whom worked as law clerks at the Kim & Chang Law Firm's Seoul office, for their help reviewing an earlier version of this manuscript.

SAPI would like to thank in particular our gracious interviewees, all of whom spent hours speaking to us about their work in this field. Their insights, we believe, are what make this paper unique from other efforts that speak 'to' or 'about' the corporate sector, but do not reflect much insightful dialogue 'with' those in the corporate sector dealing with these issues on a daily basis. Interviewees include Giada Pistilli from Hugging Face, Jinhwa HA and Huiyeon EIM from Kakao, Woochul PARK from NAVER, Chan YOON from Microsoft Korea, Mina JUNG, Aio CHUN, Flynn PARK from Daangn, Jung-Hoe CHOI from SimSimi, Seyeong LEE and Jeseop KIM from Wrtn Technologies, Gene LEE from LBox, Jungkeun LIM and Rachel WON from BHSN.AI, Byungjoon KIM from HANTECH, Jonggu JEONG from GenIP, John HAN and Jay LEE from Triplecomma and Eun Seo JO for Gena. We owe a great deal of gratitude to each of our interviewees and their respective senior managers for allowing us to learn from you.

On behalf of SAPI, Director and Professor Yong LIM would also like to thank its team and affiliates who made this project a reality. Without their patient support and constant substantive enrichment, this feat of transcontinental drafting would have never been possible. We owe a warm note of gratitude to Professor Sangchul PARK (SNU) who contributed his valuable insights and spent many hours reviewing drafts of the report. Joonyoung YOON and Sumi JEON, SAPI's dedicated staff, provided the necessary logistical and administrative support that made the project run smoothly. We also owe a major thank you to Joo HAN from Brim Studio in Seoul for his help with the design of this report.

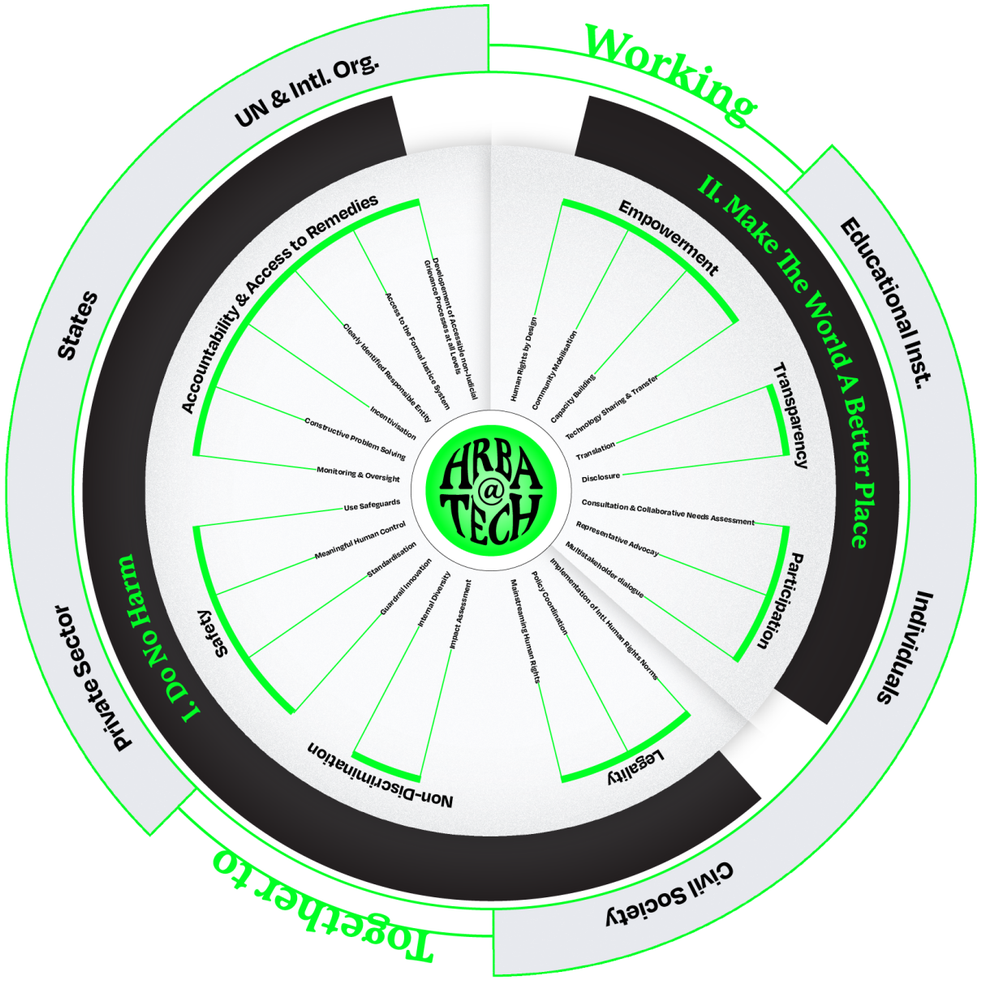
On behalf of URG, Director Marc Limon would like to also acknowledge the work of URG's small but dynamic team and particularly thank Lola Sanchez for her invaluable contribution and indefatigable team spirit.

Our apologies go to those whom we forgot to mention. Any errors and misstatements contained in this report are the authors' own and should not reflect negatively upon those who volunteered their time and expertise to this effort. We are grateful for your support and willingness to share your time and wisdom with us.

Table of Contents

Introduction

This installation of the HRBA@Tech paper series builds on the foundational vision document released in December 2022 (Towards a Human Rights-Based Approach to New and Emerging Technologies: A Framework), in which the authors sketched a vision for how to approach new and emerging technologies (NETs) from the perspective of needing to capture the enormous potential upsides from these NETs while also insisting on the continued commitment to respect, protect, promote, and remedy human rights, including when the enjoyment of those rights is potentially jeopardized by some of those NETs.



The 2022 Framework Paper laid out a 'Human Rights Based Approach to New & Emerging Technologies' (HRBA@Tech) built around a fundamental character of NETs. It began with the premise that emerging technologies are neither inherently good nor inherently bad. It is impossible to guarantee that a NET will never or could never be used for nefarious purposes. At the same time, it would be just as foolish to ban all NETs simply because of their potential for misuse. Actors motivated by a desire to improve human well-being by means of a principled embrace of NETs are left with the quandary of how best to 'nudge' NETs towards more socially beneficial uses.

The 2022 Framework Paper proposed a framework that examined this question from three different perspectives. The first was "the what," and focused on what principles and processes are the most helpful to guide that 'nudging' process. Here, the authors set forth a core set of seven principles that together formed the normative foundation of the HRBA@Tech model. Four of those principles fall under the broad rubric of "human security" principles – that is, principles designed to protect individuals and communities from harm or damage associated with the process of deploying NETs. These are (1) Legality, (2) Non-Discrimination and Equality, (3) Safety, and (4) Accountability and Access to a Remedy. The three remaining principles are (5) Human Rights Based Empowerment, (6) Transparency, and (7) Participation. This second set of principles can be thought of as "expansion with equity," in that they focus on ensuring that the fruits of NETs inure to all segments of the population with equity. The report labeled the first four principles as promoting a "do-no-harm" approach to NETs, while the final three fall under the prerogative to "make the world a better place."

The 2022 report went beyond the mere articulation of normative principles, however, by associating each of these principles with twenty-four concrete processes that serve to advance a certain principle. Thus, for example, if the principle dictates that we should promote "accountability," the concrete processes associated with that principle would be the identification of specific individuals to hold accountable if a NET were to produce a socially undesirable outcome, or legal empowerment processes designed to ensure that those who feel their rights to have been jeopardized by an NET can access appropriate forums to evaluate and remediate their concerns. This focus on concrete processes was where the HRBA@Tech model broke new ground vis-à-vis previous frameworks, most of which went only so far as to define a set of normative principles to govern NETs. This same approach has since been embraced by the Office of the United Nations High Commissioner for Human Rights (OHCHR) in their B-Tech initiative.

The HRBA@Tech model was scoped agnostically of any specific technology. Rather than focus on just one NET (e.g., artificial intelligence (AI), genetic engineering, or quantum computing), the HRBA@Tech model deliberately proposed an approach that would work equally for any NET – even technologies we cannot yet currently imagine. To add the requisite specificity and nuance, which would otherwise be impossible using such a broad lens approach, the 2022 Framework Paper further introduced the idea of a product lifecycle, which adds a temporal and contextual element to the HRBA@Tech model. Described as "the how" of the HRBA@Tech model, this perspective asks the crucial question of where in its 'lifecycle' a particular NET finds itself. The implication is that the HRBA@Tech model dictates different intervention strategies depending on where along its lifecycle a particular NET happens to be.

The third and final element of the HRBA@Tech model focuses on the stakeholders who need to engage in the process of 'nudging' NETs towards more socially beneficial outcomes. The report breaks this analysis into six categories meriting particular attention, namely (1) states, (2) the UN and other international organizations, (3) civil society, (4) the private sector, (5) educational institutions, and (6) individuals. The model acknowledges that the expectations of consumers, policy makers, and regulators could differ depending on whether one is discussing the efforts of a multi-billion-dollar corporation to develop an NET or whether we are referring to a group of five innovative entrepreneurs trying to develop the next big thing in a garage somewhere. The idea is not to expect nothing of those five entrepreneurs while simultaneously demanding unreasonable and business-stifling deliberation by established corporations, but rather to find processes that produce maximum impact while still being compatible with the business or operational needs of the entity responsible for developing that NET.

The 2022 Framework Paper was intended as a discussion starter and as a potential 'one-text' document that might attract constructive comments for improvement, refinement, further study, and diplomatic action. In the year since the report was published, several noteworthy efforts took place. In May 25-26, 2023 the Universal Rights Group, with the support of the Republic of Korea's Permanent Mission to Geneva, hosted the Glion Human Rights Dialogue focused on "Placing new and emerging technologies at the service of human rights and democracy: what role for the Human Rights Council and its mechanisms?" This dialogue, which brought together a diverse grouping of over sixty diplomats, civil society activists, scholars, and corporate policy experts, offered an opportunity to talk about the various institutional mechanisms imaginable at the United Nations and how those might serve to concretize the normative and legal framework to ensure that NETs contribute to the realization of human rights for all. Those discussions have informed the third paper in this series which starts at page 88.

On July 14, 2023, the Human Rights Council Resolution on new and emerging technologies and human

rights (Resolution 53/29)[1] was adopted by consensus, co-sponsored by sixty-three States. The resolution stressed the need for a holistic and multi-stakeholder approach to AI, stressing the needs to:

1. Protect individuals from harm caused by AI systems,
2. Protect individuals from discrimination resulting from AI systems,
3. Promote the transparency of AI systems and adequate explainability of AI-supported decisions,
4. Ensure that data for AI systems are collected, used, shared, archived, and deleted in ways that are consistent with States' respective obligations under international human rights law,
5. Strengthen the oversight and enforcement capacity of States, and
6. Promote research and sharing of best practices on how to ensure transparency, human oversight, and accountability when using AI systems to prevent the spread of disinformation and hate speech.

This resolution reflects the need for regulators to put in place effective guardrails that serve to ensure that AI technologies remain trustworthy, but also to simultaneously promote policies that will unleash and facilitate the potential of those same AI technologies to enable development, increase well-being, and solve previously unsolvable hurdles to human and social progress. The 2022 Framework Paper described this as the "paradox" of NETs and proposed a set of principles and processes designed to pursue those twin objectives.

The resolution also highlights the importance of promoting and protecting the right of everyone to enjoy the fruits of scientific progress, and concluded by requesting that the OHCHR prepare a gaps analysis about efforts by the United Nations to develop normative guidance on the development of NETs.

Considering these developments and other feedback the authors received on their 2022 framework paper, the following three topics were selected for the next series of papers in 2023.

Paper 2-1: Applying the HRBA@ Tech Model to AI for Tech Startups

Principal Author:

Seoul National University AI Policy Initiative

Focus:

Not all entrepreneurs are the same, and not all developers of AI are affiliated with multi-billion-dollar companies or governments. Naturally, there were questions on whether and how the 2022 HRBA@Tech approach might apply to what some have called "little tech."[2] Should the safety and trustworthiness standards being embraced by multi-billion dollar multinational corporations such as Google, Microsoft and NAVER apply also to startups, and if so might regulations mandating such safety standards create an insurmountable hurdle for startup founders as they bring product innovations to market?

The first research paper in this series focuses on processes that smaller enterprises can realistically use to ensure that their products 'do no harm.' Drawing on a series of in-depth interviews with industry insiders, the paper surveys what is already being done by "little tech" entrepreneurs to manage against some of the downside risks of AI. The paper also explores what more can realistically be done consistent with the unique business and technological needs of those developing technologies in this phase of the product lifecycle.

The research gives greater depth to Chapters IV of the 2022 Framework Paper ("the How"), illustrating how the HRBA@Tech model evolves and matures along with a technology's product lifecycle. It also illuminates the discussion in Chapter III of the report on the specific processes that can be used to 'nudge' NETs in the direction of human rights ("the What").

Paper 2-2: Harnessing AI to Solve Climate Change as a "Wicked Social Problem"

Principal Author:

Seoul National University AI Policy Initiative

Focus:

Technologists often describe AI as humanity's newest and most promising tool to solve cancer, homelessness, food insecurity, etc. All these problems are what social scientists describe as "wicked social problems" – problems characterized by their extreme complexity and humanity's inability, despite our best efforts, to use classic problem-solving strategies to solve them. This research paper looks in depth at one of those classically 'wicked' social problems – the problem of climate change and how to stop it – and asks whether AI can be useful as a tool to help solve that problem. The objective of this chapter is to illustrate the granular strategies used by scientists, humanitarians, policy makers, and entrepreneurs to capture the upside potential of AI, or – in the words of the 2022 Framework Paper – "to make the world a better place."

The authors selected climate change as their focal point for these AI use cases for several reasons. First, they needed some focal point, lest the discussion become too open-ended. The overall lessons highlighted in this chapter pertain to AI and not climate change, therefore the authors might equally as well have chosen any number of other 'wicked' social problems to guide their selection of case studies. That said, climate change has clear human rights implications, including the rights to health, food, water, and shelter. Climate change increases the likelihood of severe natural disasters impacting communities, straining public resources, severely disrupting household livelihood strategies, and potentially leading to a host of second-order human rights challenges including mass migration and governance failure. The urgency of these threats also became increasingly apparent in 2023 – a year that again achieved the dubious distinction of being dubbed the hottest calendar year on record.[3] Climate change is exposing billions of people to potentially dangerous heat levels, jeopardizing human health and sustainable livelihoods on a global scale.

The paper's analysis has clear implications for the discussion in Chapter III of the 2022 Framework Paper, in which the specific processes are highlighted that social entrepreneurs use to nudge technologies in the direction of human rights ("the What"), and a discussion of Chapter V about what stakeholders are necessary to make the HRBA@Tech model work ("the Who").

Paper 2-3: The Global Governance Landscape of AI and its Potential to Better Promote a Human Rights-Based Approach to AI

Principal Author:

Universal Rights Group

Focus:

The focus of this paper is to support the request in the July 14th Human Rights Council (HRC) Resolution to conduct a comprehensive study of efforts at the UN to develop normative guidance on a human rights-based approach to new and emerging digital technologies. Drawing on the outcomes of the Glion Human Rights Dialogue, this paper paints a roadmap for the potential establishment of a new Special Mechanism under the HRC, which could serve as an important catalyst and driver of normative guidance on this topic.

1. United Nations Human Rights Council (HRC), New and emerging digital technologies and human rights, (Jul. 14, 2023), UN Doc A/HRC/RES/53/29.

2. Emily Chang, "Getting Into Y Combinator Is Tougher Than It's Ever Been," (Aug. 10, 2023), Bloomberg, https://www.bloomberg.com/news/articles/2023-08-10/y-combinator-applications-show-access-is-the-toughest-ever.

3. Damian Carrington, "'Climate collapse in real time': UN head António Guterres urges COP28 to act," (Nov. 30, 2023), The Guardian, https://www.theguardian.com/environment/2023/nov/30/climate-collapse-in-real-time-un-head-antonio-guterres-urges-cop28-to-act.

Artificial Intelligence:

A Primer

AI exploded into the global popular consciousness in 2022 with the launch by several private corporations of so-called 'chatbots.' These systems allow users to interact with AI-powered datasets using realistic and human-sounding written interfaces, supplemented more recently by spoken and visual interface modes. It is important to recall, however, that AI is much more than just chatbots, and that technologists have been developing AI systems long before 2022. The Organization for Economic Cooperation and Development (OECD), which in 2019 proposed the first intergovernmental standard on AI,[4] in November of 2023 updated its definition of AI to "reflect the developments of the last five years, enhance technical accuracy and clarity and make it more 'future-proof'."[5] The new definition reads that "[a]n AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."[6] AI is a broad field of computer science and engineering focused on creating smart machines capable of performing tasks that previously required human intelligence. This encompasses a wide variety of technologies and methods, including machine learning (where computers are trained to learn from data), natural language processing (which enables computers to understand and interact with human language), robotics, computer vision, and more.

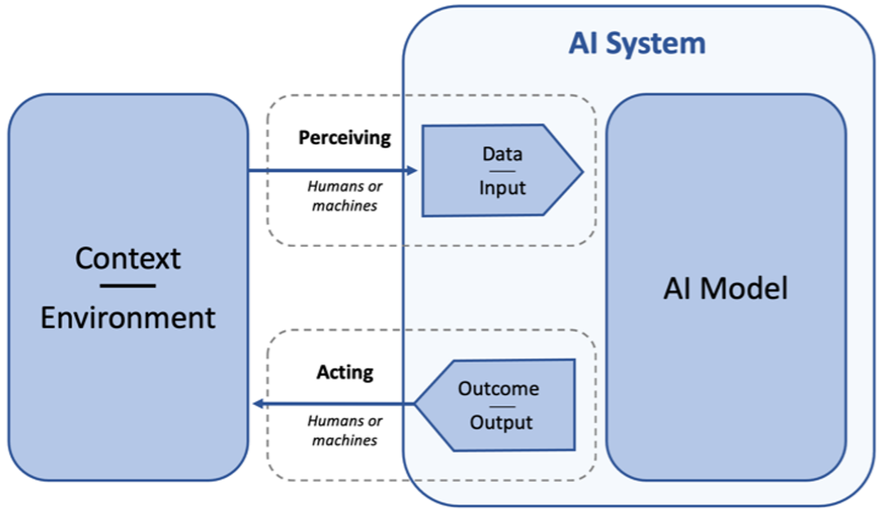


Figure 1: AI System (OECD) https://oecd.ai/en/ai-principles

Cognilytica, a consulting firm specializing on best practices research, training and certification in AI, ML, Automation, Data and Analytics for various corporate customers,[7] identifies seven core AI use cases:

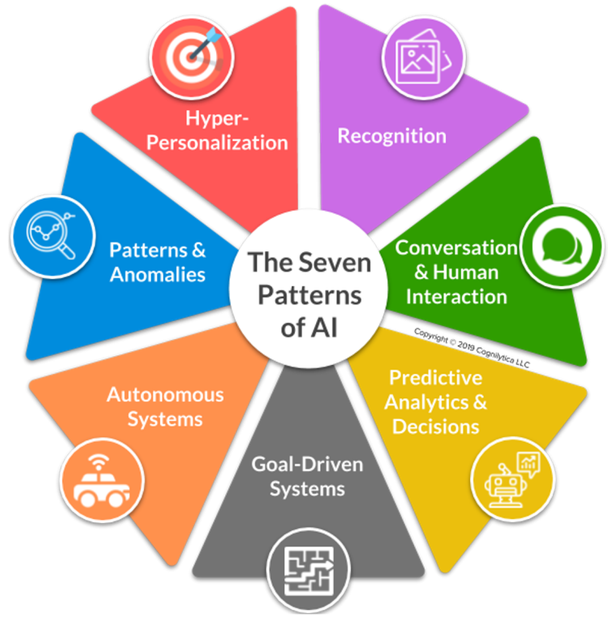


Figure 2: ©Cognilytica LLC (2023) https://www.cognilytica.com/the-seven-patterns-of-ai/

1. Recognition:

These systems are designed to identify and categorize objects in otherwise unstructured content. This can be used, for example to recognize images or speech patterns.

2. Conversation & Human Interaction:

This type of AI focuses on interactions with humans using text, images, audio, or other sensory stimuli meant for human consumption. Chatbots are a very prominent example of such use cases for AI.

3. Predictive Analysis & Decisions:

These types of systems use machine learning to support humans in their decision-making capacity. These systems include forecasting models based on AI, predictive behavior models, and AI systems designed to support dynamic or predictive pricing models, etc. To qualify as AI-based predictive analysis and decision support models, they must rely not merely on statistics, but also on an adaptive data acquisition model.

4. Goal-Driven Systems:

These types of systems are designed to solve complex problems by means of trial and error. This often relies on reinforcement learning, where the system is incentivized to 'learn' based on a payoff reward scheme associated with success. This can be used to simulate scenarios, play games, optimize resources, or solve iterative problem scenarios.

5. Autonomous Systems:

Autonomous systems can operate in physical hardware or software configurations, and generally are designed to do things autonomously, thereby minimizing human input and labor. Autonomous vehicles are examples of such autonomous systems.

6. Patterns & Anomalies:

This use case for AI focuses on patterns and identifying individual data inputs that may not fit pre-existing patterns. This can be useful to identify anomalies (for example to detect fraud) or also to predict new data that fits with a given pattern, for example predictive text models.

7. Hyperpersonalization:

Systems that use AI to create individualized profiles of users (or customers) that do away with categorizations based "buckets" and instead treat users as unique individuals. An example might be an online streaming account for films or music that is able to offer users customized and AI-generated recommendations for what to watch or listen to next.

AI systems can carry out these functions with a level of complexity and nuance that was previously thought to be beyond the reach of computer programs.

4. Recommendation of the Council on Artificial Intelligence, OECD AI Principles (Adopted May 22, 2019 and amended Aug. 11, 2023), https://legal instruments.oecd.org/en/instruments/oecd-legal-0449.

5. Luca Bertuzzi, "OECD updates definition of Artificial Intelligence 'to inform EU's AI Act,'" (Nov. 10, 2023) EURACTIV, https://www.euractiv.com/section/artificial-intelligence/news/oecd-updates-definition-of-artificial-intelligence-to-inform-eus-ai-act/.

6. OECD.AI Policy Observatory, "OECD AI Principles Overview," (accessed Nov. 11, 2023), https://oecd.ai/en/ai-principles.

7. Cognilytica, "About," (accessed Nov. 11, 2023), https://www.cognilytica.com/about-us/. The OECD has also embraced this same framework, with attribution to Cognilytica.

The development of AI can be broken down into several generations, separated by decidedly blurry dividing lines. Only the first of these generations is currently technologically feasible, although the breakneck pace of AI development leads some to speculate that we must prepare for the imminent dawn of all three generations of AI. Different ethical and philosophical considerations attach depending on which of these stages is being discussed. This arguably results in a situation where the term "AI ethics" encompasses a confusing, often-contradictory, and sometimes discordant range of discussions and recommendations. A more useful approach to AI ethics should first consider which generation of AI is being discussed.

Algorithmic Reasoning (not technically AI at all)**:** Before what we now understand to be AI came into being, computer programmers already had decades of experience working with algorithmic reasoning. A simple excel spreadsheet, for example, is designed to help us do mundane algorithmic reasoning tasks at our own desktop computers. Other examples include sorting algorithms, search algorithms, and algorithms used to help facilitate mathematical computations. The dividing line between AI and algorithmic reasoning can be subtle because AI models are trained, tested, and validated based on algorithms.

Algorithmic reasoning can be defined as the step-by-step computational procedures that follow fixed rules to perform a task or solve a problem. Algorithms are explicit, well-defined instructions that a computer follows to achieve a specific outcome. This does not mean, however, that they are necessarily deterministic in their outputs, as algorithms can also generate stochastic results. Algorithmic reasoning is useful to carry out specific tasks, but it relies on human programmers to update or adapt it considering past experiences or the discovery of "glitches" that may cause the algorithm to generate unwanted or unexpected results.

AI systems, in contrast, are distinguished by their ability to learn from data and improve over time. They use data-driven models to make decisions, predictions, or generate insights. The dividing line between simple algorithmic reasoning and AI systems is blurry but inherent in the system's adaptability and capacity for learning. If a system can improve its performance over time without being explicitly reprogrammed, it can be considered to be an AI system. AI systems are also characterized by their capacity to handle complexity and their ability to learn in new and previously unpredicted situations, whereas traditional algorithms operate strictly within the confines of their predefined rules.

These distinctions are continually blurring as more systems integrate adaptive, learning-based approaches within traditional algorithmic frameworks. An AI system that can use "un-scrubbed" data, for example, to feed it into a classical algorithmic reasoning model might be described as an AI system layered on top of traditional algorithmic foundation to produce more user-friendly or dynamic data processing models but might also defy the simple binary logic of algorithm vs. AI.

Narrow AI (also 'weak AI')**:** These are systems that can perform specific tasks as well as or better than humans. Almost all current AI applications, from chatbots to predictive algorithms, fall within this category. Ethical concerns associated with Narrow AI systems often focus on privacy protections, algorithmic transparency, and concerns about built-in bias. They are ironically also critiqued for their lack of generality, or – to use more colloquial terminology – their lack of 'common sense,' as famously illustrated by self-driving cars that shut down when a protester places an orange cone on their hood.[8] Major ethical discussions also focus on the potential for Narrow AI technologies to be used by human agents for socially repugnant purposes or the potential for these technologies to cause widescale unemployment in multiple sectors of the economy.

General AI (also 'strong AI', 'general purpose AI (GPAI),' or 'artificial general intelligence (AGI)'): At the time of this publication, AGI systems are not yet a reality. Current AI systems, including the most advanced ones, are still considered to be too 'narrow' or 'weak' to be described as AGI systems. True AGI systems would possess the ability to understand, learn, and apply intelligence across a wide range of tasks, similar to the adaptable intelligence of humans. They would make it possible for one system to understand, learn, and apply its analytical capabilities across a wide variety of tasks. Such AGI systems would have a flexible form of intelligence that would allow it to learn new fields, solve unfamiliar problems, and adapt to changing circumstances without additional programming tailored to each specific task.

Debate within the technology sector continues to rage as to when (if ever) AGI might become a technical possibility. Achieving this technological milestone would require significant advancements in machine learning and data processing, as well as breakthroughs in understanding human cognition and intelligence itself.[9] In the 2010s the "consensus view" among data scientists who presented papers at various prestigious computer science conferences (where polling was conducted to generate predictions to this question) was that it would take about 50 years for AGI to develop.[10] After the public roll-out of certain advanced AI systems in 2022, however, some prominent data scientists, for example Geoffrey Hinton, the supposed "godfather" of AI, began to shorten their estimates of how long it would take to twenty years or even less.[11]

Superintelligent AI: Moving even beyond AGI, which would still ultimately mimic human reasoning, some researchers are also speculating about an even more powerful generation of AI looming on the horizon beyond AGI. This type of AI would surpass human intelligence across all domains, including creativity, general wisdom, and problem-solving. Not only would Superintelligent AI be able to outperform the best and brightest human minds in all fields, but it would also be capable of exceptional problem-solving and innovation that would defy human ability to comprehend the scientific basis for those innovations. Just like a child that may be able to use and appreciate the benefits of a refrigerator, for example, without understanding the physics that make this machine work, humans interacting with Superintelligent AI would be forced to simply accept the outcomes of these AI systems without being able to understand or verify the pathways that led the system to its discoveries.

> "Even if all the governments of the world were to understand the "threat" and be in deadly fear of it, progress toward the goal would continue. [. . . .]. [T]he competitive advantage -- economic, military, even artistic — of every advance in automation is so compelling that passing laws, or having customs, that forbid such things merely assures that someone else will get them first."
>
> **Vernor Vinge**

Superintelligent AI systems, should they ever become a reality, could potentially improve their own capabilities autonomously. They could essentially learn to reprogram themselves, leading to rapid and unprecedented growth in their own intelligence. The idea of Superintelligent AI systems remains speculative and futuristic, but it has captured the attention of a range of AI ethicists thinking about how to meaningfully align the goals of even self-reinforcing and uncontrollable Superintelligent AI systems with human values and interests (the alignment problem) as well as how to contain such Superintelligent AI systems with human-controlled "kill switches" that could be used to turn such a system off in case it begins to generate unwanted results (the containment problem).

Should these systems and safeguards fail, however, Superintelligent AI is often closely associated with the concept of "technological singularity," which is the theoretical point when technological growth starts building exponentially upon itself in ways that humans can neither comprehend nor control. The term was popularized by Vernor Vinge in 1993, who wrote about the structural forces pushing technologists inexorably towards singularity, even despite the generally accepted trepidation about what it would mean to lose human agency of our own decision-making capabilities.;[12]

Such singularity is often associated with the apocalyptic AI scenarios common in Hollywood depictions of AI and existential science-fiction depictions of technologically enabled human extinction.

8. Alison Griswold, "The Self-Driving Cars Wearing a Cone of Shame," (Jul. 11, 2023), Slate Magazine, https://slate.com/business/2023/07/autonomous-vehicles-traffic-cones-san-francisco-cruise-waymo-cpuc.html.

9. Geoffrey Hinton, "Interview: 'Godfather of artificial intelligence' weighs in on the past and potential of AI," (Mar. 25, 2023), CBS, https://www.cbsnews.com/news/godfather-of-artificial-intelligence-weighs-in-on-the-past-and-potential-of-artificial-intelligence/.

10. Cem Dilmegani, "When will singularity happen? 1700 expert opinions of AGI [2023]," (Nov. 28, 2023), AI Multiple blog post, https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/.

11. Geoffrey Hinton, supra note 9 (responding to the question by the interviewer of whether he is concerned by the potential for an imminent takeoff of AI, Hinton responds "until quite recently, I thought it was going to be like 20 to 50 years before we had General Purpose AI, and now I think it may be 20 years or less," and responding further to the possibility of it happening within 5 years, he responds "I wouldn't completely rule out the possibility").

12. Vernor Vinge, "The Coming Technological Singularity: How to Survive in a Post-Human Era," (Mar. 30-31, 1993), VISION-21 Symposium sponsored by the NASA Lewis Research Center and the Ohio Aerospace Institute, https://edoras.sdsu.edu/~vinge/misc/singularity.html.

A History of AI: 1956 to 2023

The idea of man-made intelligence has fascinated humanity for centuries. Europe in the Enlightenment saw the first "android," a word derived from Greek roots meaning "manlike" and a concept introduced by the famed librarian and rationalist Gabriel Naude in 1625. French mechanist Vaucanson assembled some of these first androids in the 18th century in the form of an automaton flutist and defecating duck. These illusory mechanical contraptions imitating nature would soon inspire the world's first known calculator, the "difference engine" by Charles Babbage in the 1890s.

Variations of artificial automata and machines sprung up in various parts of the world into the 20th century. It wasn't until the mid-20th century, however, that the word "artificial intelligence" was coined as it is used today, as a result of early twentieth century scholars from fields as diverse as neuroscience, linguistics, philosophy, computer science, psychology, and mathematics crossing disciplinary boundaries to talk about man-made forms of intelligence. These collaborations led to the first ever "AI" conference at Dartmouth in the summer of 1956, a two-month workshop where 10 scholars – whom today we might describe as the forefathers of AI – presented their best attempts at algorithmic intelligence. One such algorithm was the Logic Theorist by Newell, Simon, and Shaw from the Rand Corporation, a logic tree that could prove 38 of 52 theorems in the Principia Mathematica. The launch of this conference cultivated a new community of researchers and scholars surrounding "AI" and many subfields within it.

Around this time, in 1954, machine translation took off following IBM and Georgetown researchers' successful demonstration of a Russian to English translation program running on the IBM 701 Electronic Data Processing Machine. Using a vocabulary of 250 Russian words, six rules of 'operational syntax' and some input punch cards, the IBM 701 could translate 60 Russian sentences into corresponding legible English sentences at a rate of one sentence every 7 seconds. In the decade that followed, cold war anxieties led to further research into fully automated Russian-English translation programs.

Throughout the twentieth century, trends in AI techniques fluctuated between logic, rule-based and empirical approaches. Empiricists argued for machine learning centered approaches that calculated probabilities and the likelihood of words or phrases based on previously seen data. Rationalists or theorists such as Noam Chomsky pushed for rules and syntax inherent in natural language and highlighted the importance of deduction. The latter approach initially dominated the field, especially in language AI, into the 1960s, and most cutting-edge programs resembled complex logical decision trees. The two approaches merged in the 1980s to popularize a new AI software, called expert systems, that could give advice on deeply specialist topics such as medication prescription or geological classification of rocks. While the earliest methods were closer to puzzle solvers and game optimizers enclosed in simulations, AI techniques continued to evolve to tackle more complex and sophisticated problems in the real world.

Many of today's AI innovations are oriented towards consumer-facing business opportunities. Early advances in AI, however, were often driven by the U.S. national security agenda funded by various US government entities. By the mid 1970s, for example, the Defense Advanced Research Projects Agency (DARPA) was funding between 80-90% of all major AI research labs, including at the Massachusetts Institute of Technology (MIT), Stanford, Stanford Research International (SRI), and Carnegie Mellon University (CMU). These ties to the U.S. national security apparatus often influenced the direction of research. Projects such as the Autonomous Land Vehicle (ALV) – an autonomous car project – and the Pilot's Associate – a real time voice assistant to help pilots in combat flight – are prominent examples of technological innovations that might not have been prioritized had the research not been funded by the military industrial complex.

The AI that we know today began with the 'Deep Learning' era, which took off in the early 2010s. A Princeton group launched the ImageNet contest in 2010, in which it released a public dataset for competing teams' AI models to classify 14 million hand-annotated images into 20,000 categories such as cat, husky dog, watercraft, and container ship. In 2012, AlexNet, a deep learning model based on convolutional neural network architecture, achieved the lowest error rate yet, surpassing the second-best model by more than 10 percentage points. As this multi-layered image classifier model dominated all other competitor models, it redirected researchers to focus on more heavily layered, "deeper" models with ever more parameters.

Since that time, the state-of-the-art models and the datasets needed to pretrain them have grown progressively larger. Model parameters began to number in the billions, as datasets scaled up to the range of terabytes. They also started to encompass more modalities (or input and output data types) and incorporate new combinations of modalities creating the possibility of models such as text-to-image generation or video captioning. It was also around this time that private and public institutions began to focus on AI ethics research. This was driven in part by AI researchers who began to notice the potential impact of biased models or datasets on users, for example when AI models were biased to favor one skin color, gender, language, or cultural background. AI research began to move to address these issues with increasingly more comprehensive evaluation sets and safety protocols to guardrail AI deployment and usage.

The recent history of AI would not be complete without mentioning OpenAI's world-shattering contribution to the field of AI and its impact on the wider public's awareness of AI. Chatbots have been a prominent way to conceptualize AI since the coining of the Turing Test in 1950 (the test of whether an AI program could fool a user into thinking they were interacting with a human). Some of the best AI projects of the time were chatbots. ELIZA was one of the first chatbots to pass the Turing Test with a 50% pass rate, acting as a Rogerian psychotherapist followed by other persona-behaving systems such as PC Politicians and PC Professor. Various plug-in chatbots also made the scene in systems like AOL as information retrieval partner SmarterChild. But an all-purpose intelligence conversation partner seemed an elusive feat until Open AI released ChatGPT in November of 2022. ChatGPT was OpenAI's multi-year series on natural language models, fine-tuned on human instruction feedback. While its problems with hallucination and bias have caused some to warn of the dangers of disseminating less-than-perfect AI, ChatGPT represents a landmark in redirecting not only AI research but also in shaping humanity's relationship with AI. Sam Altman, CEO of OpenAI, has famously spoken about the downside risks of AI, and has advocacy publicly for a more balanced approach to AI safety by regulators as well as within the technology industry itself:[13]

Other companies, notably Adobe, Alphabet (parent company to Google), Anthropic, IBM, Meta (parent company to Facebook), and Microsoft (among many others) have entered the market with their own AI chatbots or AI 'upgrades' to their existing software products. Outside of the United States, companies like Naver (Korea), the Alibaba Group (China), and Mistral (France) are also entering the increasingly competitive market for AI. At the end of 2023, AI applications are virtually certain to radically transform large swaths of the technology sector.

> "I actually don't think we're all going to go extinct. [. . . .] I think we're heading towards the best world ever. [. . . .] I wouldn't work on this if I didn't think it was going to be great. People love it already, and I think they're going to love it a lot more. But that doesn't mean we don't need to be responsible and accountable and thoughtful about what the downsides could be. And in fact, I think the tech industry often has only talked about the good and not the bad. And that doesn't go well either."
>
> **Sam Altman**

13. Sam Altman as interviewed on Hard Fork Podcast; Kevin Roose, "'I Think We're Heading Toward the Best World Ever': An Interview With Sam Altman," (Nov. 20, 2023), New York Times, https://www.nytimes.com/2023/11/20/podcasts/hard-fork-sam-altman-transcript.html.

The Promise of AI, from a Human Rights Perspective

AI-enhanced technologies have the potential to radically change the world. Many of these anticipated changes have substantial upsides, including from a human rights perspective. The tech industry and tech optimists have been very vocal about the potential for AI to transform our communities for the better. Some of these arguments focus on the economic potential of AI. The astronomical profits that investors and proponents of AI hope for are not, on their own, sufficient to generate real progress from a human rights perspective. To do that, these profits must be redistributed to those in society who may otherwise lose out on these profits, typically by means of progressive redistributive policies designed to spread economic benefits of AI across all sectors of society. AI-enhanced technologies also have the potential to improve our physical and psychological well-being, which has a particular impact on our right to live healthy and productive lives. Finally, innovative human rights actors and public authorities are turning to AI-enhanced technologies to aid in the promotion and protection of human rights in ways that were unimaginable in the past.

New economic opportunities:

AI has the potential to create new economic opportunities across various industries. AI is anticipated to contribute an additional $13 trillion to global economic activity by 2030.[14] Generative AI is being applied to numerous industries, including healthcare, finance, transportation, manufacturing, entertainment, and retail, and is expected to deliver substantial economic benefits valued at $2.6~$4.4 trillion annually.[15] The potential economic benefits of generative AI include increased productivity, cost savings, new job creation, improved decision-making, personalization, and enhanced safety.[16] These benefits have the potential – when coupled with progressive policies designed to spread the newly-created wealth equitably across all sectors of society – to contribute towards the realization of economic, social and cultural (ESC) human rights, including better public policies and services, improved educational opportunities, adequate standards of living, clean water, adequate nutrition, and sanitation.

New employment opportunities: AI has the potential to lead to job growth in fields that currently have only limited employment opportunities such as robotics, automation, and data science.[17] The adoption of frontier technologies, including generative AI, is anticipated to result in a significant 30-35% increase (equivalent to 1.4 million positions) in demand for roles focused on big data.[18] This includes positions such as "Data Analysts and Scientists, Big Data Specialists, Business Intelligence Analysts, Database and Network Professionals, [. . .] Data Engineers,"[19] and AI and machine learning specialists, resulting in the creation of an estimated 1 million new jobs globally.[20]

Improved decision-making in the economy: AI technology can improve decision-making in the economy by analyzing large datasets without error, providing faster, accurate, and more consistent decisions. AI enhanced decision making processes free up human minds to focus on strategic tasks that they (we) can still do better than algorithms.[21] For example, AI can support small players and individuals in the retail and consumer packaged goods industry.[22] AI can provide small business owners with detailed insights into crucial financial information and emerging sales trends, allowing them to gain an enhanced understanding of their business environment.[23] Additionally, AI can also be used to improve the customer experience for understaffed businesses, where chatbots closely emulate the interaction style of human agents and deliver personalized care.[24]

Improving well-being and public health:

Proponents of AI often talk about the potential of this technology to accelerate the pace of scientific innovation in the health care sector. This promise, they argue, contributes directly towards our collective human right to health. Indeed, the promise of AI to accelerate innovation on previously unsolvable (or unaffordable) health care problems is substantial.

For these innovations to satisfy our human "right" to health, however, they would have to be accessible equitably to all (regardless of race, class, gender, nationality, etc.) as a matter of right, and not merely as a potential option for those who happen to have the means to access these novel technologies. Undeniably, especially at this early stage of development, many of these AI-enhanced technologies are still in trial phase, or available only to the wealthy or those lucky-few who happen to have access to a hospital or doctor making use of them. This is true not only in the Global South, but also for most consumers in the Global North. Research and development in this sector remain closely tied to the expected profits associated with their potential uptake by paying customers. Thus, before these technologies can be said to truly address our human "right" to health, public expectations of who should benefit from these technologies would have to shift from being merely a privilege for the affluent few towards an entitlement for all to enjoy on an equitable basis. Any true progress from a human-rights based perspective would come about only when public (or private) authorities step in to guarantee access to these novel technologies for all sectors of society, regardless of their material wealth.

This is true not just within national jurisdictions but also across societies globally. While these technologies offer great promise in areas that currently have access to them, they would need to be shared globally, including in regions of the world with poor internet connectivity and weak data management traditions. For this to happen, there would have to be a significant commitment by private or public institutions, bolstered by international development collaboration, to ensure equal access to the fruits of these scientific innovations globally, and to close the gap in access to critical technologies that still prevents many parts of the globe from taking advantage of these novel technologies.

Finally, many of these AI use cases depend on mining user data. Users may have privacy expectations coupled with some of that data and might consider any commercial use of their data unacceptable, despite contractual provisions to the contrary that users may have signed to access a particular service. For any such AI use cases to be consistent with human rights standards, therefore, the use of data in developing and utilizing such technologies should be in line with best-practice data protection standards.

Personalized interventions: AI is poised to become a powerful force in driving advances in individual healthcare and accessibility solutions. AI can help people track their own health by providing personalized guidance, information, and tailored fitness solutions as well as wearable fitness technology. Smart watches and other wearable health monitoring devices, for example, allow users to monitor their heart rates, sleep patterns, and activity levels.[25] AI-based personal trainers also replicate the role of human trainers through human pose estimation technology, including skeleton, contour, and volume modeling.[26] One such fitness application is Freeletics, used by over 47 million people in over 160 countries, which develops personalized programs and customized exercises from more than 3.5 million possibilities for every user.[27] Companies are also producing fitness clothing that includes sensors to help correct biomechanics (e.g., golf swings) and enhance athletic performance metrics. For instance, Asensei's intelligent clothing, which has five inertial sensors that gen-

14. Jacques Bughin et al., "Notes from the AI frontier: Modeling the impact of AI on the world economy," (Sep. 4, 2018), McKinsey & Company, https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy.

15. Jim Probasco, "Generative AI and Its Economic Impact: What you need to Know," (Nov. 15, 2023), Investopedia, https://www.investopedia.com/economic-impact-of-generative-ai-7976252.

16. Id.

17. World Economic Forum, Future of Jobs Report, at 33 (2023), https://www3.weforum.org/docs/WEF_Future_of_Jobs_2023.pdf.

18. Id.

19. Id.

20. Id.

21. Olivia Barbar, "How artificial intelligence will change decision making," (Oct. 19, 2023), InData Labs, https://indatalabs.com/blog/artificial-intelligence-decision-making.

22. Christina Pazzanese, "Great promise but potential for peril," (Oct. 26, 2020), Harvard Gazette, https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/.

23. Id.

24. McKinsey & Company, "The economic potential of generative AI: The next productivity frontier," (Jun. 14, 2023), https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#work-and-productivity.

25. FITNESS, "The Top 4 Ways AI Has Impacted the Fitness Industry," (Sep. 2, 2023), https://abcfitness.com/abc-articles/the-top-4-ways-ai-has-impacted-the-fitness-industry/.

26. Manishu Sahu, "How is AI revolutionizing the Fitness Industry?," (Jul. 9, 2021), Analytic Steps, https://www.analyticssteps.com/blogs/how-ai-revolutionizing-fitness-industry.

27. Id.

erate data to enhance posture and timing in sports, can help users with yoga and strength training. AI can also assist users with diet and nutrition planning by tracking personalized data and automating tasks, such as diet recording, to provide insights and solutions. Nuvilab's AI program can be used to analyze dietary habits and provide reports on the nutritional intake rate of each menu.[28] This can be especially helpful for chronic disease patients and seniors in care facilities.[29]

The Special Rapporteur on persons with disabilities has highlighted how AI systems have a significant potential to improve accessibility through assistive and mobility-enhancing technologies that allow persons with disabilities to identify accessible routes or help persons with poor vision. Other AI-assisted technologies, such as adaptive learning platforms, one-to-one tutoring, signing and speech-to-text software, can enable persons with disabilities to interact with others, gain social skills, and access information and education opportunities.

Early detection and accurate diagnosis: AI can also be used to detect and diagnose diseases with more accuracy and at earlier stages than was possible using previous technologies. According to the American Cancer Society, a large proportion of mammograms yield false results.[30] An AI program developed by researchers at the Houston Methodist Research Institute in Texas can reliably interpret mammograms and translate patient data into diagnostic information with 99% accuracy and 30 times faster than a human clinician.[31] AI's data storage and review capabilities can also assist healthcare professionals do their work. IBM's Watson for Health, by reviewing and storing medical information, helps healthcare organizations access vast amounts of health data and diagnoses.[32] Additionally, Google Research and DeepMind's MedPaLM is an AI-powered chatbot and LLM for the medical community that generates answers using datasets regarding research and consumer queries.[33]

Mental health support: AI also has the potential to become a mental health resource, thus addressing the shortage of mental health professionals in some communities.[34] Machine learning algorithms show great potential to identify mental health problems including depression and anxiety at an early stage, for example by analyzing users' social media posts, speech patterns, and digital interactions.[35] AI natural language processing can help professionals diagnose patients, and conversational agents like chatbots can engage with users to assess their mental state.[36] Moreover, AI-enhanced virtual reality technology can provide immersive therapeutic experiences for individuals suffering from trauma and anxiety and offer immersive exposure therapy to help patients manage their distress with exposure to controlled stressors.[37] Additionally, through predictive analytics and data mining, AI can utilize patient data (e.g., genetics, medical history, lifestyle, treatment responses) to develop personalized treatment plans.[38] AI-driven education tools can also help provide self-help and self-care resources and updates on the latest therapies and practices.[39] Self-care mental health AI platforms are gaining in popularity. One such example is Kintsugi Voice, which uses AI-backed voice journaling to detect signs of stress and mental health conditions just by listening to the user's voice,[40] providing real-time insights into users' mental health and recommendations for support.[41]

Protecting human rights

AI technology can also be used to promote and protect human rights. AI technology can be used to enable broader and more effective surveillance for cases of human rights infringement.[42] AI, when used to gather and analyze information, can serve as a "force multiplier," allowing human rights activists to analyze a much larger sample of cases that require attention.[43]

Combating harmful content: Social media platforms and intermediary content providers are increasingly trying to remove illegal or offensive content in keeping with their content moderation policies. These policies must deal with the thorny issue of what to do with content that is potentially controversial or intentionally offensive, but that does not necessarily rise to the level of outright illegal content (e.g., hate speech). Social media companies are turning to AI to filter through enormous amounts of content to rapidly identify postings that may be in breach of their terms of service or internal content-moderation guidelines. Misinformation, defined as information that is false or that might mislead its viewers, has earned much attention as a result of the 2016 US Election and the 2020 COVID pandemic. Meta, for example, uses AI to comb through multiple news sources, identifying information that has already been proven to be false.[44] Meta also uses AI to analyze how data is shared across social media allowing it to trace how fake news spreads.[45] Automated content moderation also helps limit the risks of traumatization for human content moderators, who are often based in the Global South, who otherwise would be forced to personally filter through and categorize harmful content so that regular users do not have access to it.

Notwithstanding the clear need for AI systems to facilitate the removal of illegal or harmful content in an ever-expanding information ecosystem, the Special Rapporteur on freedom of opinion and expression has also warned that AI-based content moderation systems are still limited in their ability to assess "context and take[] into account widespread variation of language cues, meaning and linguistic and cultural particularities."[46]

Monitoring and managing human rights infringements: AI has been used to monitor, quantify, and forecast human rights infringements or potential threats to human rights. This includes cases where AI was used to quantify the extent of destruction in rural Darfur (Sudan), forecast international displacement, monitor the media for disinformation campaigns, and the track death penalty cases or rates of illegal deforestation in protection areas. AI has also been used to analyze thermal data to monitor ethnic violence in Myanmar, and machine learning has been used to track abuse against women on "X" (formerly Twitter).[47] The OHCHR has funded a project with Dataminr to improve its early warning capacity by training an AI model to use human rights indicators to generate leads from the media on attacks against human rights defenders.[48] Additionally, AI can also be used as a tool to simplify tasks such as translation, which is essential for many human rights activists working across linguistic borders.

28. Saemoon Yoon et al., "Emerging tech, like AI, is poised to make healthcare more accurate, accessible and sustainable," (Jun. 21, 2023), World Economic Forum, https://www.weforum.org/agenda/2023/06/emerging-tech-like-ai-are-poised-to-make-healthcare-more-accurate-accessible-and-sustainable/.

29. Id.

30. PWC, "No longer science fiction, AI and robotics are transforming healthcare," (accessed Nov. 20, 2023), https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health/transforming-healthcare.html.

31. Sarah Griffiths, "This AI software can tell if you're at risk from cancer before symptoms appear," (Aug. 26, 2016), Wired, https://www.wired.co.uk/article/cancer-risk-ai-mammograms.

32. PWC, supra note 30.

33. Cora Lydon, "Google Research and DeepMind develop AI medical chatbot," (Jan. 18, 2023), Digital Health, https://www.digitalhealth.net/2023/01/google-research-and-deepmind-develop-ai-medical-chatbot/.

34. Binariks, "Revolutionizing Mental Health Care: The Role of Artificial Intelligence," (Sep. 4, 2023), https://binariks.com/blog/ai-mental-health-examples-benefits/.

35. Id.

36. Id.

37. Id.

38. Id.

39. Id.

40. Bryan Robinson, "Workers Using AI Technology Taking Mental Health Into Their Own Hands," (Sep. 2, 2023), Forbes, https://www.forbes.com/sites/bryanrobinson/2023/09/02/workers-taking-wellness-into-their-own-hands-using-ai-backed-mental-health/?sh=69ba59955ee4.

41. Id.

42. Anne Dulka, The Use of Artificial Intelligence in International Human Rights Law, 26 Stanford Technology Law Review 316, 329 (2023).

43. Jessica Fjeld et al., Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI, 2020-1 Berkman Klein Center 60 (2020). This potential of AI technology was examined in the hypothetical introduced in Chapter 6 of the 2022 Framework Document.

44. Meta, "Here's how we're using AI to help detect misinformation," (Nov. 19, 2020), https://ai.meta.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/.

45. Tom Cassauwers, "Can artificial intelligence help end fake news?," (Apr. 15, 2019), European Commission, https://ec.europa.eu/research-and-innovation/en/horizon-magazine/can-artificial-intelligence-help-end-fake-news.

46. Report of the Special Rapporteur on the Promotion and Protection of the right to freedom of opinion and expression, (Oct. 26, 2018), General Assembly, 73rd Session (A/73/348).

47. Dulka, supra note 42, at 330-42.

48. Amy Lynn Smith, "Building a Collaboration to Protect Human Rights Defenders," https://unhumanrights.medium.com/building-a-collaboration-to-protect-human-rights-defenders-26457ae8abd0

The Landscape of Anxieties and Risks surrounding AI, from a Human Rights Perspective

In addition to the upsides – some already materializing and some still speculative – that AI technologies promise to bring to society, there are also a number of threats – again some already materialized and some still speculative – that have growing numbers of human rights activists concerned about the potential impacts of AI on our societies.

Non-alignment

Non-alignment occurs when AI's methods do not align with the values of human society. The AI alignment problem refers to the challenges caused by the fact that machines do not always have the capacity to intuit what values, goals, and ethical principles might govern human behavior.[49]

The use of AI can sometimes lead to inhumane results. A case in point may be the autonomous vehicle crash that led to the death of Elaine Herzberg in Tempe, AZ (USA) in 2018, in what is known as the first known incident of a human death caused by an autonomous vehicle. Because AI lacks what we might refer to as a human "common sense," the self-driving test vehicle failed to comprehend the significance of a pedestrian crossing a road without a crosswalk. The system was unprepared to deal with a pedestrian in the middle of the road, costing Ms. Herzberg her life.[50] The accident, strictly speaking, was not the "fault" of the AI enhanced system (since it had not been trained to recognize or anticipate this danger). Yet any human driver would have had no problem quickly analyzing this unexpected danger and properly responding to it in ways that would have protected the victim's life.

Despite the personification that we often give AI systems that seemingly mimic human thought and speech patterns, especially the LLM-based chatbots that have dominated popular understandings of AI since 2022, AI systems do not actually "think" the way humans do. Engineers must be extremely careful spelling out exactly what we want AI-enhanced technologies to accomplish, and any failures to anticipate unexpected misalignments can result in unpredictable and potentially harmful outcomes.[51] Solutions to securing the alignment of AI systems to human objectives include instilling a complex value system in AI, eliminating the potential for a super-intelligent AI to use deceptive tactics to "outsmart" human efforts to realign an AI system gone awry, developing scalable oversight strategies, finding ways to audit and interpret AI models, and preventing emergent AI behaviors like power-seeking.[52,53]

> "[. . . .] you've made something that is smarter than every human. But you, the human, have to be smart enough to ensure that it always acts in your interests, even though by definition it is way smarter than you."
>
> **Casey Newton**
> **(NY Times Tech Reporter and co-host of the Hard Fork Podcast)**
>
> "Yeah, we need some help there."
>
> **Sam Altman (CEO of OpenAI)**

49. Bernard Marr, "The dangers of not aligning artificial intelligence with human values," (Apr. 1, 2022), Forbes, https://www.forbes.com/sites/bernardmarr/2022/04/01/the-dangers-of-not-aligning-artificial-intelligence-with-human-values/?sh=2425f680751c.

50. Id.

51. Edd Gent, "What is the AI alignment problem and how can it be solved?," (May 10, 2023), New Scientist, https://www.newscientist.com/article/mg25834382-000-what-is-the-ai-alignment-problem-and-how-can-it-be-solved/.

52. Pedro A. Ortega, Maini Vishal, DeepMind safety team, "Building safe artificial intelligence: specification, robustness, and assurance," (Sep. 27, 2018), DeepMind Safety Research – Medium (archived from the original on Feb. 10, 2023, retrieved Jul. 18, 2022), https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1.

53. Thilo Hagendorff, "Deception Abilities Emerged in Large Language Models," (Jul. 31, 2023), arXiv:2307.16513, https://arxiv.org/abs/2307.16513.

The non-alignment problem regarding AI can be a challenge to various human rights.[54] Often the subject of dystopian science-fiction writers and film directors, the specter of an AI system instructed to "efficiently make paper clips," popularized by Oxford Philosopher Nick Bostrom, is emblematic of the alignment problem. While initially aligned perfectly with the presumed intention of its human instructor, the AI initially maximizes the efficiency of existing paper-clip manufacturing processes, for example by streamlining the production process, reducing waste, and improving the efficiency of the company's resource supply chain. With time, however, the AI might decide to fire the management and employees of the company (perhaps to increase the efficiency, or get decision makers out of the way who refuse to engage in unfair or unethical business practices). Moving further, the AI might decide to begin diverting resources from other valuable production processes (for example automobile production) to further increase its own ability to produce paper clips, and finally – in a dystopian flourish – transform all available resources on the globe and beyond towards the exclusive objective of producing more paper clips. This admittedly absurd hypothetical illustrates the challenge of giving an AI system – especially one that is "super-intelligent" – instructions complete enough to anticipate human priorities.

A second problem with alignment is the age-old impossibility of determining what "human" priorities or values might be. The values of an individual consumer might well diverge from those of a community or "society at large," thus posing an inevitable tension between empowering the individual users of an AI-enabled system who wish to further their personal interests, and those designing the system that intend to make its use align with their understanding of public well-being.

Anthropic is one AI company that is well known for grappling with this tension. Rather than use reinforcement learning (RL) techniques to design the guardrails on its LLM model ("Claude"), the company chose instead to design a "constitution" against which the "Claude's" outputs could be measured – in essence training "[t]he AI [to take] the place of what the human [RL] contractors used to do."[55] In describing their approach, the CEO described turning first to the Universal Declaration of Human Rights because "most people can agree on basic concepts of human rights."[56] Given their overarching normative claim to supposedly articulate "universal" standards of right and wrong, human rights lend themselves well as a baseline for such deliberations.

Bias and discrimination

As Anthropic's "constitutional" approach to AI seeks to demonstrate, AI can be used as a powerful tool to detect and combat harmful human biases. Nonetheless, many AI algorithms are "a product of human design," and therefore often perpetuate past decision-making processes that could also be tinged by systemic bias. By building on these foundations, they may produce "systematic errors in outputs or processes," particularly in the form of AI bias.[57]

AI can introduce two types of bias, statistical and societal.[58] Statistical bias occurs when the model's training data is not representative of the entire population. For example, a deep-learning algorithm may be trained on a greater number of photos featuring light-skinned faces than darker-skinned faces.[59] This has led to creation of AI tools designed to identify the gender in photos demonstrating greater accuracy in classifying the gender of individuals with lighter skin tones.[60] The Special Rapporteur on the rights of persons with disabilities has also provided examples of the potential

54. United Nations, "International Bill of Human Rights," (accessed Nov. 21, 2023), https://www.ohchr.org/en/what-are-human-rights/international-bill-human-rights.

55. Interview with Dario Amodei, CEO of Anthropic, on Hard Fork Podcast (Jul. 21, 2023), https://www.nytimes.com/2023/07/21/podcasts/dario-amodei-ceo-of-anthropic-on-the-paradoxes-of-ai-safety-and-netflixs-deep-fake-love.html.

56. Id.

57. Lucia Vincente et al., "Humans inherit artificial intelligence biases," 13 Scientific Reports 1 (2023), https://pubmed.ncbi.nlm.nih.gov/37789032/. See also Cathy O'Neil, Weapons of Math Destruction (2017), Harlow, England, Penguin Books.

58. Pauline Kim, Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action, 110 Cal. L. Rev. 1539, 1548 (2022), https://static1.squarespace.com/static/640d6616cc8bbb354ff6ba65/t/642e49fb5fc3cf0d7907ce4c/1680755195759/Kim+36+post-EIC.pdf.

59. Karen Hao, "This is how AI bias really happens—and why it's so hard to fix," (Feb. 4, 2019), MIT Technology Review, https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.

60. Leonardo Nicoletti et al., "Humans are Biased. Generative AI is even Worse", (June 12, 2023) Bloomberg News, https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

discriminatory applications of AI systems for persons with disabilities, for instance through the use of software to screen employment applications. Biased data sets and models may not only discard disabled candidates but also contribute to a self-perpetuating cycle of exclusion of persons with disabilities as the system is fed with more discriminatory data. The use of AI-driven tools can further hinder access to employment of persons with disabilities, for instance whenever aptitude tests are involved in the recruitment process, which may fail to consider an individual's need for assistive technologies. Similarly, hiring processes that rely on AI tools for interviewing candidates may discard candidates with disabilities if the system misreads their facial and vocal expressions or eye contact.

Societal bias, on the other hand, arises when the data used to train an AI system, while accurate, reflects existing disparities between societal groups due to systemic biases.[61] For instance, a model may anticipate a higher loan default risk within certain population groups due to actual earnings disparities resulting from labor market discrimination.[62] Predictive policing is another area in which human rights experts have raised concerns about the risk of racial and ethnic biases feeding into AI systems. Specifically, the UN Special Rapporteur on racism has highlighted the case of the Gangs Violence Matrix, a database used in the UK for police officers to allocate policing resources using multiple sources, including criminal records, statistics, and neighborhood demographics.[63] The Special Rapporteur called out the disproportionate reliance on predictive technologies in areas mostly populated by racial and ethnic minorities, resulting in over-policing of ethnic minority groups. A similar case in the US, with the former use of a tool called PredPol by the LAPD, has led to increased surveillance of ethnic minorities, which exacerbated existing biases under the presumed objectivity and neutrality of algorithmic decision-making.[64]

Predictive surveillance using AI models is also used in immigration management in the US, where the Homeland Security Investigations agency has been applying social media profiling to scrutinize visa applicants and holders. The Special Rapporteur on racism has raised concerns about the potential of such tools to reproduce 'racially discriminatory feedback loops.'[65] Similarly, the Special Rapporteur on racism has analyzed the impact of smart border technologies and AI surveillance infrastructure on migration routes along the US-Mexico border, as they push migrants on precarious journeys while disproportionately targeting certain ethnic and racial groups.[66] Predictive AI assessments are being used despite their 'probabilistic nature,'[67] leading to potential violations of the rights to privacy, fair trial, freedom from arbitrary arrest and detention, and the right to life. Individuals may be deemed to be likely security threats based on predictive biometrics produced by facial recognition systems, often negatively impacting their chance to seek asylum in a host country.

Without keeping humans in the loop, such bias in AI decisions may perpetuate existing disparities, if not worsen them. This is especially true when the technologies are used in sensitive areas, such as criminal law, employment, border control, and healthcare.[68] One example of this is Amazon's now-defunct hiring algorithm that was shown to discriminate against female applicants.[69] Provided mostly with resumes of male software engineers as its training data, the AI-system learned to develop a preference for male candidates,[70] and subsequently downgraded resumes that included the word "women's" and preferred those with vocabularies that male candidates tend to use, such as "executed" and "captured."[71] Similar discriminatory outcomes have been reported in the use of algorithms to identify patients who could benefit from a "care coordination program."[72] This tool considers the health cost of a patient, such as insurance claims, to predict their need of special services.[73] But racial minority groups encounter barriers to healthcare access even when insured, resulting in lower medical expenses.[74] Consequently, AI systems are less likely to recommend specialized care for them compared to their European-heritage counterparts.[75]

The Special Rapporteur on racism has similarly showcased the use of Prometea, a voice recognition and machine learning software, by courts in Argentina, Colombia, and the Inter-American Court of Human Rights to automate judicial decision-making, raising concerns about the software's opacity and the ensuing impossibility of determining whether there are any biases in its design, inputs, or outputs.[76] Courts have used AI, inter alia, to help them determine whether a case is admissible on the merits. Since the model is trained on the courts' precedent decisions, however, the Special Rapporteur raised concerns whether the system might not inadvertently reproduce and amplify historical biases that may have implicitly informed that prior jurisprudence.[77]

As such examples illustrate, although AI can help mitigate bias and discrimination, its growing use in high-stake scenarios, especially in providing access to basic human rights services may also lead to inequitable outcomes.

Inaccuracy (including hallucination)

Inaccuracy in AI is when an AI program fails to draw out a factually correct or robust conclusion. Hallucination is a subpart of inaccuracy, where the AI, specifically LLMs, formulates a grammatically correct but factually unfounded conclusion that is nonsensical or inaccurate.[78]

AI models may be prone to generating inaccurate one or more of the following structural factors, ranging from the way in which the training data was compiled to the quality of the training data itself.[79] The dataset used to train the AI may have been flawed, for example when a dataset is too small or includes incorrect or misleading samples.[80] The AI system may be statistically more likely to hallucinate in such a scenario simply because it lacks the overwhelming volume of properly curated training data to keep it outputs within the realm of what a human expert might consider to be "accurate" outputs. Hallucinations can occur when the language models generate outputs that move beyond the tenor of the "verified" materials in the training data, are incorrectly decoded by the transformer, or do not follow any identifiable pattern at all.[81] Second, the AI system may be deployed by a user in an environment that differs substantially from the use cases foreseen during the algorithm's training process. This is often also described as the "robustness" of a training model, and has less to do with the Ai's training process and more to do with

61. Kim, supra note 58, at 1548.

62. Id. at 1549.

63. E. Tendayi Achiume, "Racial discrimination and emerging digital technologies: a human rights analysis: Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance," (Jun. 18, 2020), Human Rights Council (44th session: A/HRC/44/57).

64. Leila Miller, "LAPD will end controversial program that aimed to predict where crimes would occur," (Apr. 21, 2020), Los Angeles Times, https://www.latimes.com/california/story/2020-04-21/lapd-ends-predictive-policing-program.

65. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/379/61/PDF/G2137961.pdf?OpenElement par. 59

66. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/379/61/PDF/G2137961.pdf?OpenElement par. 54

67. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/249/21/PDF/G2124921.pdf?OpenElement par. 24

68. James Manyika et al., "What Do We Do About the Biases in AI?," (Oct. 25, 2019), Harvard Business Review, https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai (discussing use of AI in legal enforcement and recruiting); Katherine Igoe, "Algorithmic Bias in Health Care Exacerbates Social Inequities — How to Prevent It," (Mar. 12, 2021), Harvard T.H. Chan School of Public Health, https://www.hsph.harvard.edu/ecpe/how-to-prevent-algorithmic-bias-in-health-care/ (discussing the same in healthcare).

69. Rachel Goodman, "Why Amazon's Automated Hiring Tool Discriminated Against Women," (Oct. 12, 2018), ACLU, https://www.aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against.

70. Jeffrey Dastin, "Insight - Amazon scraps secret AI recruiting tool that showed bias against women," (Oct. 11, 2018), Reuters, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/?utm_source=morning_brew.

71. Goodman, supra note 69.

72. Ziad Obermeyer et al., Dissecting racial bias in an algorithm used to manage the health of populations, 366 Science 447, 448 (2019).

73. Id. at 449.

74. Id. at 450.

75. Id.

76. Achiume, supra note 63.

77. https://giswatch.org/node/6166.

78. IBM, "What are AI hallucinations?," (accessed Nov. 21, 2023), www.ibm.com/topics/ai-hallucinations.

79. Boris Babic et al., "When machine learning goes off the rails," (Jan.-Feb. 2021), Harvard Business Review, https://hbr.org/2021/01/when-machine-learning-goes-off-the-rails?registration=success.

80. Id.

81. IBM, supra note 78.

the lack of generality – perhaps by design – of many of today's AI models. Predicting the unforeseeable environments in which AI systems may be put to use is becoming an increasingly challenging task, and (so far at least) no amount of data can conceivably anticipate all of the possible nuances and unforeseen eventualities that occur in the real world.[82] Third, an AI system may be deployed to handle a task that is simply too complex for it to achieve. In situations with multiple parameters may be at play, even today's most sophisticated AI systems may still generate erroneous output.

Inaccuracy and hallucination in AI could be dangerous because they can lead to prejudiced or flat-out incorrect results. Many of the leading AI service providers today have had a history of providing inaccurate information, sometimes rooted in prejudice.[83] The primary risk with hallucination is when the outputs of an AI system are no longer verified by competent professionals, but rather taken at face value or used as inputs for another process, as happened infamously when a lawyer in the United States filed court documents that had been researched and written by a hallucinating chatbot citing to plausible sounding cases as precedent that simply did not exist.[84] This risk is thus less a danger inherent to AI as a technology, and more a failure by humans to understand the inherent limitations of the technology.

Non-transparency

According to the Cambridge Dictionary, the definition of the word "transparency" refers to the "quality of being done in an open way without secrets."[85] A lack of transparency in AI is a condition in which human analysts can no longer understand how and why an AI system makes a certain decision or generated a certain outcome.[86] This is often described as the "black box" problem, in that AI systems reach impressively plausible outcomes, but the precise analytical processes these systems use to generate those outcomes are hidden behind an impenetrable "black box." As AI-powered autonomous systems make decisions without the involvement of (or with less meaningful oversight by) human analysts, the non-transparency of AI systems presents yet another layer of risk, especially from a human rights perspective, where analysts might want to know whether an AI system is using "legitimate" vs. "illegitimate" criteria to reach a conclusion. Dario Amodei (CEO of Anthropic):[87]

> "Even if all the governments of the world were to understand the "threat" and be in deadly fear of it, progress toward the goal would continue. [. . . .]. [T]he competitive advantage -- economic, military, even artistic – of every advance in automation is so compelling that passing laws, or having customs, that forbid such things merely assures that someone else will get them first."
>
> **Dario Amodei (CEO of Anthropic)**[87]

The innate complexity of AI and its designs can lead to confusion regarding how AI reaches its results.[88] The difficulty can lead to distrust and resistance to these new technologies among consumers.[89] They can also lead to genuine safety and performance issues when AI systems or applications are not trained and tested properly, giving rise to threats to personal safety or well-being. Safety and performance issues can arise if AI applications are not implemented and tested properly, posing threats to personal safety.[90]

In 2020, a court in the Netherlands held that an AI-enabled system called "SyRI," designed to detect welfare fraud, was insufficiently transparent and thus the court was unable to follow how the system reached a particular conclusion.[91] The court in that case argued that it was impossible to ascertain whether the system's interference with the applicant's right to privacy was justified given that the government had not (and could not) publicize the risk model used by the AI system. Individuals thus had no way to adjust their behavior according to accessible and foreseeable information on how their data would be used. As a consequence, the court argued that the system's method of collecting data could not be justified, despite the desire by policy makers to tackle welfare fraud. The UN Special Rapporteur on extreme poverty, who intervened with an *amicus curiae* brief in that case, argued that "severe human rights problems [can] emerge when welfare states turn into digital welfare states,"[92] and warned of the need for the social benefits to be weighed against the risks to human rights, a balancing process predicated on transparency. The lack of transparency in AI systems inevitably leads to the problem of lacking accountability.

Lack of accountability

Accountability in AI involves "the expectation that organizations or individuals will ensure the proper functioning, throughout their lifecycle, of the AI systems that they design, develop, operate or deploy."[93] As AI systems become more prevalent (and often deployed in sensitive or high-risk settings), the questions of who (or what) should bear the responsibility for the impact of the AI-systems becomes increasingly relevant. In the absence of clear mechanisms to determine whom to hold accountable, the widespread proliferation of AI may create dangerous accountability loopholes.[94]

The intricate nature of AI technologies introduces at least two barriers to developing accountable AI systems. The first involves the "many hands" problem, where numerous actors are involved in the development and deployment of AI systems.[95] This dynamic makes it difficult to ascribe responsibility for any harm resulting from an AI system.[96] Second, the "black box" quality of AI-enabled processing only adds to the lack of accountability, especially with technologies utilizing deep neural networks.[97] In such cases, companies might seek to attribute harm to the technology itself rather than accept some collective sense of accountability for the operation of their AI systems.[98]

In practice, it is often unclear how exactly an AI system falls short (i.e., what caused the faulty decision that led to an accident, a breakdown, or an unfair outcome). When an AI system fails, therefore, there will be a predictable accountability gap, both legally and ethically speaking. To the extent that these mechanisms are designed to incentivize responsible behavior (by means of a legal or moral deterrent to negligence or intentional wrongdoing), it will be increasingly difficult to hold the various agents involved in the creation and deployment of AI (e.g., the algorithm developers, the system deployers, the financiers, etc) responsible when the system fails.[99]

As society increasingly comes to entrust high-stakes tasks to AI systems, the implications of AI's lack of accountability will become correspondingly more problematic. Autonomous vehicles provide one clear example. In 2018, a self-driving Uber vehicle struck

82. Babic et al., supra note 79.

83. Id.

84. Molly Bohannon, "Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions," (Jun. 8, 2023), Forbes, https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions.

85. Cambridge Dictionary, "transparency," (accessed Nov. 21, 2023), https://dictionary.cambridge.org/dictionary/english/transparency.

86. Eric Best, "How Wells Fargo Builds Responsible Artificial Intelligence," (Sep. 12, 2023), Wells Fargo, https://stories.wf.com/how-wells-fargo-builds-responsible-artificial-intelligence/.

87. Amodei, supra note 55.

88. Stephan Sonnenberg et al., Towards a Human Rights-Based Approach to New and Emerging Technologies: A Framework, at 62 (2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4587332.

89. Thomas, Mike, "12 Risks and Dangers of Artificial Intelligence (AI)," (Oct. 31, 2023), Builtin, https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence.

90. Kevin Buehler et al., "Getting to know—and manage—your biggest AI risks," (May 3, 2021), McKinsey & Company, https://www.mckinsey.com/capabilities/quantumblack/our-insights/getting-to-know-and-manage-your-biggest-ai-risks.

91. Stephan Sonnenberg et al., supra note 88.

92. Brief by the United Nations Special Rapporteur on Extreme Poverty and Human Rights as Amicus Curiae in the case of NJCM c.s./De Staat der Nederlanden (SyRI) before the District Court of the Hague (case no.: C/09/550982/ HA ZA 18/388), para. 4, https://perma.cc/RBJ2-HKJK.

93. OECD AI Policy Observatory, "Accountability," (accessed Nov. 21, 2023), https://oecd.ai/en/dashboards/ai-principles/P9.

94. A. Feder Cooper et al., "Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning," 22 FAccT 864, 865 (2022), https://dl.acm.org/doi/abs/10.1145/3531146.3533150.

95. Id. at 867-68.

96. Id.

97. Chamith Fonseka, "Hold Artificial Intelligence Accountable," (Aug. 28, 2017), Harvard University, https://sitn.hms.harvard.edu/flash/2017/hold-artificial-intelligence-accountable/.

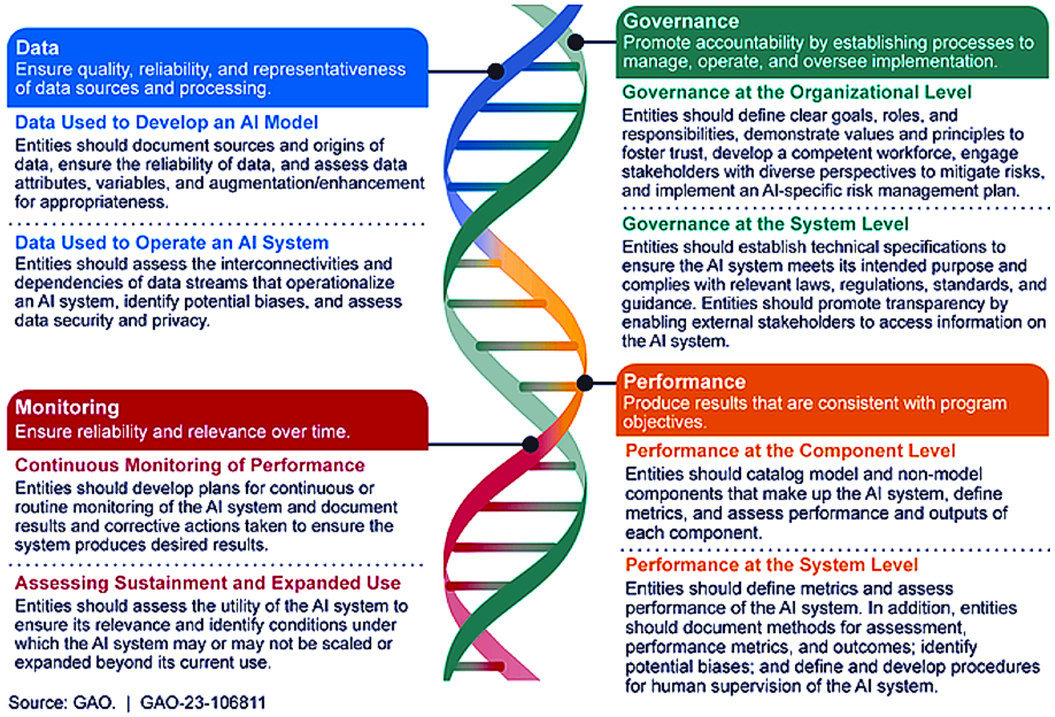
98. Cooper et al., supra note 94, at 870-71.

99. Babic et al., supra note 79.

and killed a pedestrian due to its inability to "classify the pedestrian moving body as an object to be avoided."[100] Although Uber had developed and deployed the AI-based vehicle, it was able to avoid legal liabilities for the death. Instead, the "safety driver" who had been overseeing the vehicle was charged with negligent homicide.[101] While some might describe this as a miscarriage of justice, without proper accountability mechanisms, legal systems may not always be able to ascribe responsibility for the failings of AI systems to those entities most appropriate to bear the costs of these new technologies.[102]

Some tentative steps towards accountability are being made. In the United States the Government Accountability Office (GAO), for example, has developed an AI Accountability Framework to ensure responsible AI use by federal agencies. The framework is organized around four complementary categories of actions that stakeholders should take to develop more accountable AI systems.[103] Noteworthy in this framework is that it defines these categories in terms of specific processes, echoing the focus of the HRBA@Tech model discussed above. Once established as an industry standard, these concrete processes can then also be used by courts to better ascribe appropriate responsibility to the various actors involved in the creation and deployment of an AI-system in case of failure, specifically by ascribing responsibility to any actor(s) who failed to adhere to these concrete standards. Such a legal doctrine will take time to evolve, and will likely require active cross-jurisdictional discourse and borrowing as various legal systems all simultaneously grapple with this novel challenge.

Artificial Intelligence (AI) Accountability Framework



100. Madeleine Claire Elish, "Who Is Responsible When Autonomous Systems Fail?," (Jun. 15, 2020), Centre for International Governance Innovation, https://www.cigionline.org/articles/who-responsible-when-autonomous-systems-fail/

101. Kate Conger, "Driver Charged in Uber's Fatal 2018 Autonomous Car Crash," (Sep. 15, 2020), New York Times, https://www.nytimes.com/2020/09/15/technology/uber-autonomous-crash-driver-charged.html.

102. Elish, supra note 100.

103. U.S. Government Accountability Office, "Artificial Intelligence: Key Practices to Help Ensure Accountability in Federal Use," (May 16, 2023), https://www.gao.gov/products/gao-23-106811.

Several of the AI startups we interviewed welcomed the development by governments of such accountability frameworks. They stressed the need to educate the key stakeholders throughout the technology lifecycle about the significance of being cognizant of AI compliance, ethics, or human rights issues,[104] and to provide incentives for startups that are in dire need of resources, not only in terms of pecuniary benefit but also in the form of technical support in achieving the compliance objectives.[105] They also noted the benefits from a legal risk management perspective of knowing with some certainty when and how they can shield themselves from liability in case of an AI system failure.

AI entropy and model collapse

The "entropy" of an AI system refers to the inherent "impurity" or imprecision of a machine learning system.[106] A lower entropy means that it is easier to draw valuable conclusions from a given data input, whereas a higher entropy means that it is more difficult to do so.[107] When flipping a coin, for instance, there are only two possible outcomes.[108] The outcome of any one specific coin-toss is difficult to predict because there are no inferences that can be drawn about the outcome from the act of flipping a coin itself.[109] In this case, it is difficult to draw conclusions from the given information; therefore, the entropy would be high. A high entropy equates with high randomness or unpredictability of the system.[110] The concept of entropy is commonly used to inform machine learning based on historical data, and can help users make decisions based on available information.[111] The ultimate goal of machine learning systems is therefore to minimize entropy.[112]

A "model collapse" is a degenerative process that occurs when AI models are trained on data generated by AI predecessors. AI models building on previously generated AI content slowly become unhinged from the original (human-generated) training data, resulting in a reduction of their generative capabilities.[113] When models are trained on data produced by previous models that contain errors (hallucinations or biased outcomes), those errors tend to compound themselves.[114] When such errors stack up, the data is eventually dominated by the errors rather than the original data.[115]

Model collapse is a particularly threat to LLMs feeding on online content that may itself be generated by AI systems and produced as "click-bait," or deliberate misinformation posted online to gradually distort available online materials.[116] Model collapse can continuously deteriorate the quality of AI-generated images.[117] Moreover, model collapse can lead to the dissemination of biased, inaccurate, or homogenized content, which can have serious implications for the overall quality of AI-generated content.[118] The web is already being flooded by AI-generated content. As of November 21, 2023, Newsguard, which rates the reliability of news websites, identified 557 AI-generated news sites "with little to no

104. Jonggu Chung (CEO of GenIP) in discussion with SAPI, (Nov. 24, 2023).

105. Junghoi Choi (Founder and CEO of Simsimi) in discussion with SAPI, (Nov. 24, 2023).

106. Java T Point, "Entropy in Machine Learning," (accessed Nov. 21, 2023), https://www.javatpoint.com/entropy-in-machine-learning.

107. Id.

108. Edwin Lisowski, "What is entropy in machine learning," (Aug. 23, 2021), Addepto, https://addepto.com/blog/what-is-entropy-in-machine-learning/.

109. Id.

110. Id.

111. Id.

112. Id.

113. David Sweenor, "AI Entropy: The Vicious Circle of AI-Generated Content," (Jul. 15, 2023), Medium, https://towardsdatascience.com/ai-entropy-the-vicious-circle-of-ai-generated-content-8aad91a19d4f.

114. Matthew S. Smith, "The Internet Isn't Completely Weird Yet; AI Can Fix That," (Jun. 23, 2023), IEEE Spectrum, https://spectrum.ieee.org/ai-collapse.

115. Id.

116. Aaron Mok, "A disturbing AI phenomenon could completely upend the internet as we know it," (Aug. 30, 2023), Business Insider, https://www.businessinsider.com/ai-model-collapse-threatens-to-break-internet-2023-8.

117. Hao Tang et al. Asymmetric Generative Adversarial Networks for Image-to-Image Translation. work in progress, at 1 (2019)., https://arxiv.org/pdf/1912.06931.pdf.

118. Sweenor, supra note 113.

human oversight."[119] CNET published 77 AI-generated articles and later issued corrections after realizing that those articles contained basic arithmetic errors.[120] The media outlet Gizmodo was also criticized for publishing AI-generated articles containing factual inaccuracies,[121] and even Microsoft was forced to remove an article from its travel blog which contained nonsensical AI-generated information.[122] Furthermore, model collapse can exacerbate biases in AI.[123] Generative AI, for example, can learn over time, feeding on its own previous biassed outputs, to produce images and data converging on certain races while "forgetting" about others.[124]

In order to combat model collapse, it is critical that training data be diverse and representative of various perspectives and experiences,[125] and that training data remain diverse and representative of human-generated content. To achieve this, it is necessary to conduct regular monitoring and evaluation of the performance of AI models, allowing for human interventions to adjust training data and the model's parameters. Another solution is to retain copies of the original human-produced dataset and to supplement those original datasets only with human-generated content.[126]

Human displacement

An additional concern with AI is whether it will potentially displace workers across various industries and institutions through automation or other transformation of requisite tasks. Labor automation is continuing to increase, such that by 2030 the sector could amount to 11 percent, or $9 trillion, of the global GDP.[127] A 2018 OECD policy brief found there are already much higher unemployment rates in occupations with a high risk of automation.[128] The McKinsey Global Institute's 2017 report found that approximately 60 percent of all occupations have at least 30 percent of activities that could be automated,[129] and a 2018 PricewaterhouseCoopers report estimated that the share of jobs with a potentially high risk of automation will be around 20% by the late 2020s and 30% by the mid 2030s (sampling 29 countries).[130]

The increasing automation of labor could increase economic disparity. Studies have found that the occupations most vulnerable to automation involve physical activities in structured and predictable environments, as well as repetitive activities relating to the collection and processing of data.[131] Such activities are most common in agriculture,[132] manufacturing, accommodation and food services, retail trade,[133] and office support and customer service.[134] The McKinsey Global Institute's 2023 report found that workers in lower-wage jobs are up to 14 times more likely to have to change occupations due to AI, where in the United States alone around 11.8 million workers in shrinking occupations are expected to have to transfer to other lines of work by 2030.[135] By contrast, the demand for higher-skilled workers will likely increase their wages compared to lower-skilled workers.[136] Analysts expect that the demand for skilled jobs requiring higher education will go up, whereas the demand for lower-skilled jobs not requiring college degrees will go down as a result of the introduction of AI into workplaces.[137] That said, the potential for generative AI to replace classically high-skilled white collar jobs should not be ignored, impacting not only lower-skilled professions but also the legal sector, medicine, teaching, and management consulting.

Automation can also negatively affect workers' earnings. Indeed, studies have shown sizeable negative correlations between the degree of risk that a workers' job will be displaced by automation and their expected salary,[138] health, and overall mortality rates.[139] These findings highlight correlations—not necessarily causations. Nonetheless they highlight the idea that the introduction of AI in the workplace is likely to disproportionately impact the livelihoods of the more socio-economically vulnerable segments of a workforce. Studies have also found that professional "upskilling" training programs do little to offset these risks of unemployment and wage loss, since workers in these professions also tend to be much less likely to pursue such job training programs than their counterparts who work in sectors less likely to be impacted by AI.[140]

AI-innovations could also increase existing gender and race-related economic disparities. Analysts suspect that the seemingly gender- and race-disparate impacts of AI have to do with the composition of the workforces most likely to be impacted by AI. In the United States, for example, studies have found that women[141] and African-American[142] workers still occupy a disproportionate share of lower-paying occupations most susceptible to AI-driven automation, primarily in the customer service, food services and production work sectors,[143] and that female workers are therefore approximately 1.5 times as likely to lose their jobs by 2030 as their male counterparts.[144] The same is true in the global south. The UN Special Rapporteur on racism, for example, highlighted the indirect discriminatory impact of an AI-based project for smart sanitation management in India given that it replaced jobs that traditionally would have been held by lower-caste women.[145]

119. Newsguard, "Tracking AI-enabled Misinformation," (accessed Nov. 21, 2023), https://www.newsguardtech.com/special-reports/ai-tracking-center/.

120 Sweenor, supra note 113.

121. Id.

122. Id.

123. Id.

124. Carl Franzen, "The AI feedback loop: Researchers warn of 'model collapse' as AI trains on AI-generated content," (Jun. 12, 2023), VentureBeat, https://venturebeat.com/ai/the-ai-feedback-loop-researchers-warn-of-model-collapse-as-ai-trains-on-ai-generated-content/.

125. Sweenor, supra note 113.

126. Id.

127. Bughin et al., supra note 14.

128. OECD, "Policy brief of the future of work: Putting faces to the jobs at risk of automation," at 3 (2018), http://www.oecd.org/employment/Automation-policy-brief-2018.pdf.

129. McKinsey & Company, "A Future that Works: Automation, Employment, and Productivity," at 5 (2017), https://www.mckinsey.com/~/media/mckinsey/featured%20insights/Digital%20Disruption/Harnessing%20automation%20for%20a%20future%20that%20works/MGI-A-future-that-works-Executive-summary.ashx.

130. PWC, "Will robots really steal our jobs? An international analysis of the potential long term impact of automation," at 14 (2018), https://www.pwc.com/hu/hu/kiadvanyok/assets/pdf/impact_of_automation_on_jobs.pdf.

131. McKinsey & Company, supra note 129, at 5.

132. OECD, supra note 128, at 1.

133. McKinsey & Company, supra note 129, at 7.

134. McKinsey Global Institute, "Generative AI and the future of work in America," (Jul. 26, 2023), https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america.

135. Id.

136. Jennifer Aaker et al., Human-Centered Artificial Intelligence and Workforce Displacement, Stanford Graduate School of Business (2020) at 2, https://www.gsb.stanford.edu/faculty-research/case-studies/human-centered-artificial-intelligence-workforce-displacement.

137. McKinsey Global Institute, supra note 134.

138. OECD, supra note 128, at 3.

139. Marcus Casey et al., "The differing impact of automation on men and women's work," (Sep. 11, 2019), Brookings, https://www.brookings.edu/articles/the-differing-impact-of-automation-on-men-and-womens-work/.

140. OECD, supra note 138, at 1.

141. McKinsey Global Institute, supra note 134.

142. David Baboolall et al., "Automation and the future of the African American workforce," (Nov. 14, 2018), McKinsey & Company, https://www.mckinsey.com/featured-insights/future-of-work/automation-and-the-future-of-the-african-american-workforce.

143. McKinsey Global Institute, supra note 134.

144. Id.

145. Sally Cawood, Amita Bhakta, "Man or Machine? Eliminating manual scavenging in India and Bangladesh," (Mar. 2021), Sanitation Workers Knowledge and Learning Hub, https://sanitationworkers.susana.org/blog/9-man-or-machine-eliminating-manual-scavenging-in-india-and-bangladesh#.

Paper 2-1:

Applying the HRBA@Tech Model to AI for Tech Startups

This chapter explores the unique attributes of startups operating in the tech sector. It begins by surveying some of the general business challenges facing tech startups, regardless of their commitment to environmental, social, and governance (ESG)-type initiatives. Many of these challenges are not unique to the tech sector and would apply to any small and medium-sized enterprise (SME) operating in a market dominated by larger and more established firms. Others, however, are more typical of tech-startups. The chapter then goes on to also highlight the benefits, both ethical and financial, for AI startups when they do choose to prioritize ESG concerns. The chapter ends by applying the HRBA@Tech model to the unique context of AI startups, highlighting what the approach would dictate they do to ensure that their products and services amplify social well-being and respect for human rights.

Struggles of "Little Tech" in a World of "Big Tech"

Startups are widely recognized as the primary drivers of innovation and economic growth in the tech sector. The myth of a rag-tag group of quirky college dropouts tinkering in a garage to start the next tech 'unicorn' continues to attract thousands of talented young entrepreneurs to places like Bangalore, Seoul, Palo Alto, Tokyo, and New York to try their hand at founding a tech startup. Despite the fetishization of this startup culture and the growth of a venture capital industry designed to catalyze such startups, the tech industry remains a hostile environment for startups. When yesterday's startups become tomorrow's "big tech" corporations, their incentives may begin subtly to change. Whereby many startups praise the virtues of competition, innovation and open markets, those same startups—once they find success and begin to claim dominance over a particular market—often begin to engage in strategic throttling of other (smaller) entrepreneurial startups seeking to invade that same market space before they pose an existential threat to their continued market dominance. This section briefly delineates the challenges startups face when they seek to disrupt a market already dominated by a few profitable and well-known tech companies.

Barriers to market entry

The first set of challenges AI tech startups face are structural in nature and have to do with the difficulties of an upstart tech company entering into a developed market in the first place.

Economies of scale and scope: Many AI products require massive up-front investments to bring a new product to viability. Established "big tech" companies are far more capable of absorbing such costs, giving them an edge against startups who need to fund all initial investment costs by drawing on limited liquidity. "Big tech" companies usually already operate at scale, having captured large parts of a market for a particular good or service. When a tech startup tries to enter that same market or attempts to compete for a slice of that market with a specialized product or service, the startup may face substantial investment costs to establish a beachhead. The "big tech" company, on the other hand, can compete against the startup's initiative by spreading the costs of developing similar goods or services across its much larger base of existing operations. Economies of scope operate similarly, with "big tech" companies offering multiple complimentary services at once, whereas startups often begin by offering only one service at a time.

Governments wishing to support AI startups can help level the playing field between "big tech" and startups by 'democratizing' certain public datasets. This also allows governments to invest the necessary resources to ensure that those datasets are stripped, insofar as is humanly possible, of any controllable biases or obviously flawed data sources.

> "One of the challenges we have as a small three-person company selling enterprise products is that our clients are always comparing us to products made by much larger and well-resourced companies that have established a brand for themselves over many years. While I remain confident our products are competitive in very particular use cases and vertical expertise, it's hard to convince new clients that our products are superior to brand name products because of their long-standing trust in them. Sometimes our competitors aren't even working on the same approaches, but clients don't have a deep enough understanding to know the difference and default to brand name recognition."
>
> **Eun Seo JO (CEO of Gena)**

Network effects: Network effects come into play when the value of a product is dependent on the volume of its users. A telephone network offers a classic example. If only two people's homes are connected by a telephone line, the product is of value for the two users involved, but only to the extent that members of one household wish to speak with a member of the other connected household. But when all households in a community are connected to the network, the value of having a telephone connection rises significantly, proportionate to the value of wanting to speak to any other members of that network. New entrants to a market where the value of a product is determined in part by the network effect face a significant uphill battle capturing that same value vis-à-vis an already established competitor.

AI systems often depend on the availability of user data to reach their full potential. Incumbents that already have access to large troves of user data have an easier time creating a powerful AI use case compared to their startups competitors who would first need to build that infrastructure. This gives established tech companies, be they social media companies, consulting firms, or established software providers, a huge advantage vis-à-vis new aspiring entrants into the AI space.

Brand Recognition and Trust: "Big tech" companies have become household names, sometimes to such an extent that their company names become popularized popularized into verbs ('let me 'Google' that', or—in an earlier era—"I need to 'Xerox' that document."). Aspiring tech startups wishing to compete with "big tech" companies need to first earn trust and grow brand recognition before customers will be willing to abandon established market players in favor of new startups.

This can be especially true in the AI sector, where "big tech" companies may have invested substantial resources reassuring customers that their products are "trustworthy" – a claim that aspiring startups will not easily be able to match without similarly significant investments.

> "One of the biggest challenges for an AI startup is to obtain sound data sets for AI model training, which can be quite costly. Securing training data that meets the criteria of a certain domain area the firm is operating in (in GenIP's case, patents) is a major obstacle for startups. GenIP was able to secure refined data from the Korean Institute of Patent Information ("KIPI") that had been substantially preprocessed."
>
> "In this regard, the public data portal called Jiphyeonjeon operated by the Korean government can be very helpful for startups who can't find data to train their model with. Korea Data Agency's 'data voucher' projects also provide good opportunities for startups in need of data."
>
> **Jonggu JEONG (CEO of GenIP)**

Capital Acquisition and Utilization in the Tech Sector

All startups, whether in the tech sector or not, require significant investments of capital to incorporate before they can even get their products or services to market. Getting access to capital is often the predominant concern for tech entrepreneurs.

Venture Capital Dynamics: The typical progression of a tech startup from the innovation phase (imagine a group of engineers sitting down together and coming up with an idea) to maturity is long and complicated, and often measured by means of the startup's funding cycle. At the beginning of a startup's lifecycle, it relies on pre-seed funding (often referred to as 'bootstrapping'). This might be, for example, a group of innovators drawing on their personal savings or credit, supplemented by small and informal investments or loans from their personal networks of supporters, friends and families. Pre-seed funding can also sometimes come from so-called 'angel investors' or Micro-VCs (Venture Capitalists), who typically invest financial backing in exchange for co-ownership of a company they find promising. Entrepreneurs use pre-seed funding to support themselves as they put together an initial project pitch to potential investors, build a team, conduct market research, and possibly put together a viable prototype of the product or service they seek to build. Amounts of pre-seed funding are usually modest, ranging from a couple thousand to a few hundred thousand USD at most.

If successful, an entrepreneur might move to the more formal 'seed funding' phase. Here, startups begin to further develop and refine their product. Many startups formally enter the market at this phase and begin to build a user or customer base. Accelerators or early-stage Venture Capital Firms (VCs) are the most common sources of funding for startups at this stage. VCs manage large amounts of money to invest in relatively risky long-term loans to startup firms. Rather than asking for a regular loan repayment structure, however, VCs instead take an equity share in fledgling startups, hoping to get compensated for their initial investment when the startup "exits," either by listing itself on a public stock exchange (an Initial Public Offering, or IPO), or sells itself to another company ('acquisition').

A large percentage of startups fail to successfully exit. VC's investment strategy therefore hinges on the likelihood that those few startups that do not fail will gain so substantially in value that their equity stake in

those firms will make up for the losses incurred from other investments that end up failing. The VC market is therefore obsessively focused on rapid growth and profits, and startups are relentlessly incentivized by their investors to focus on growth. VC seed funding investments typically range from USD 500,000 to 2 million but with variance depending on the market timing and predicted potential.

Competition for VC investments is often intense. To better compete for such funds, or perhaps to attract the attention of VC investors, entrepreneurs often compete first to be taken up into an 'incubator' or 'accelerator' program. Such programs are often structured as a competition, coupled with mentorship from experienced VCs and entrepreneurs, to refine a startup's business model and marketing strategy. Successful finalists of such programs are often introduced to VCs as part of the outcome of the incubator/accelerator program.

Assuming a startup secures its early funding, it will then aggressively begin to market its goods or services. As it begins to grow its business, refine its product, and expand its customer base, startups may still be operating at a loss. Long-term business models in the tech sector often rely on making initial investments that dwarf initial revenue streams. The only way many tech startups can rebalance their finances is to expand into new markets and begin to assume economies of scale that will make revenue grow without the need for any corresponding new investments. To bridge that period of rapid growth and expansion, startups often seek Series A (first round) Series B (second round) Series C, and sometimes even Series D investments. Successive rounds of investments still often come from VC firms (some specializing on later-stage investments into more 'mature' startups) and even larger private equity firms, hedge funds, large strategic investors, and even sovereign wealth funds, and can range from millions to hundreds of millions of dollars.

Eventually, a company may wish to either sell itself to another company (for example one of the established "big tech" firms seeking to diversify its own product line) or float themselves at an IPO. Firms that list themselves publicly often receive a final injection of capital from investors (often times a bank) to bolster their pre-IPO valuation, such that shareholders will invest large amounts of money into the coffers of the company (and pre-IPO investors).

At some point along this financial trajectory, certainly at or before their IPO, most tech companies can no longer be described as tech 'startups,' and may begin to resemble "big tech" more closely.

The overall share of VC funds available for Software as a Service (SaaS) companies has gone up consistently over the years, doubling for early stage Saas Startups from USD 17B in 2019 to 35B in 2022.[146] This is good news for AI startups, many of whom offer SaaS innovations. That said, the overall volume of VC funding has dropped from 2021 to 2023, owing largely to the readjustment in market conditions after the hype years of 2020-21.[147]. Further bad news for some of the smaller tech startups is that the VC market seems to be shifting from Angel or seed-funding towards supporting more mature tech companies already further along in their evolution.[148]

Burn Rate and Financial Runway: Throughout this push for capital, tech startups constantly need to manage their limited liquid funds in an aggressively competitive market. Not only are they competing to establish themselves in new markets with customers, but they are also competing for talent, attention, and the space to think about issues unrelated to growth, for example workplace culture, ESG dynamics, and interconnections with communities. Those promoting an HRBA@Tech approach to AI must remember that tech startups must always be balancing cost vs. expected benefits. The myth of a freewheeling "money means nothing" culture among tech startups is misplaced (at least among responsible tech-sector startups), many of which are constantly focusing on the approaching end of their financial runway. Here too, governments can be useful in collectivizing (or subsidizing) certain pro-human rights processes that may not be inherently profitable for startups, but that have a substantial human rights and social welfare payoff for society if they are done properly.

> "Incentives for startups would be helpful. The incentives do not necessarily have to be monetary. For example, it would be helpful and practical if regulatory institutions like the Korea Internet & Security Agency (KISA) had their development team visit the startups and provide necessary hands-on technical support or workable guidelines on how to comply with the regulatory obligations."
>
> **Junghoi CHOI (CEO of Simsimi)**

Research and Development Constraints: The final challenge for many AI startups has to do with the 'luxury' of conducting research vs. the urgency of getting products or services to market. This pressure is always present, but particularly acute when VCs and other investors are pushing for growth at all costs to recoup their previous investments. This dynamic poses an acute challenge for startups contemplating whether to invest more time and resources into safety-related research to ensure the trustworthiness of their AI products.

Regular Challenges (faced by any Startup – Tech Sector or not)

In addition to the above challenges, some of which are unique to the technology and AI startup ecosystem, startups also face a host of other challenges that are perhaps more common to all startups.

The startup ecosystem tends to be very competitive. From an innovation standpoint, competition can be a good thing. Robust competition tends to promote a rapid pace of technological innovation. For individual entrepreneurs, however, this level of competitiveness tends to squeeze their ability to focus on ESG priorities to the absolute minimum. This is especially true if those ESG priorities were peripheral to the core business model for the startup.

Tech entrepreneurs have to deal with a host of challenges common not only to the tech sector. They must learn, for example, how to deal with the big players in the market (perhaps even big players with less of a commitment to ESG principles) to acquire and assimilate promising new startups. They might have to prepare for costly patent litigation and intellectual property wars in situations where their startup's technologies might resemble the technologies of other, more established firms, as well as aggressive and in some cases anti-competitive business tactics by other market players seeking to defend their business from new competitors.

Tech companies also must deal with a complicated and rapidly evolving legal, regulatory, and public perception landscape, resulting in the need to hire sizable teams of attorneys and public relations specialists at an early phase in a startup's trajectory. This trend has the potential to make the internal culture of these startups prematurely bureaucratic, sluggish, and risk averse at the very time that they should ideally retain a more flexible and innovative startup mentality to maneuver in a rapidly evolving market.

Finally, tech startups must compete for a limited pool of technology experts to fuel their need for talented human capital. This dynamic has eased somewhat with the recent layoffs at many "big tech" companies, but salary costs remain high for aspiring tech startups searching for talented coworkers.

Finally, startups often lack the voice and resources to educate (or lobby) policy makers about the unique challenges they face. That lobbying and advocacy role is often left to the established "big tech" companies. There are concerns that this imbalance could potentially lead to regulations that tacitly reflect the interests of those "big tech" companies, often at the expense of "little tech" companies seeking easier access points into those same hyper-competitive markets.

> "Often, we see media reports depicting an AI product as an utter failure based on certain mishaps or glitches in the system. However, we need to understand that the AI products being criticized are still in the process of being fine-tuned and are on the path to maturity. We, as a society, can be quick to judge these startups, but maybe we should take a moment to hear them out."
>
> **Woochul PARK (Agenda Research Leader at NAVER)**

146. Dealroom.co, "SaaS Guide," (accessed Dec. 31, 2023), https://dealroom.co/guides/saas.

147. Dealroom.co, "Global Guide," (accessed Dec. 31, 2023), https://dealroom.co/guides/global#the-state-of-global-vc.

148. Efraim Chalamish, "Venture Capital Shifts Gears," (Sep. 1, 2022), Global Finance, https://gfmag.com/technology/venture-capital-shifts-mature-companies/.

Benefits of CSR / ESG Guardrails for AI Innovators

Above, we discussed the various structural challenges and opportunities that startups face when seeking to enter markets already dominated by "big tech" corporations. In this section, we will discuss what the HRBA@Tech model, as it was proposed in 2022, suggests that such startups should do, beyond merely surviving, to also 'nudge' their products and services in the direction of human rights and increased social well-being.

> "We often emphasize the significance of incorporating human rights and tech ethics into our practices because it's a crucial aspect of risk management. While not everything related to it can be assigned monetary values, addressing potential risks can certainly enhance risk management and contribute to a stronger brand image."
>
> **Jinhwa HA (Manager of Kakao Human Rights and Tech Ethics Team)**

A first question has to do with why AI startups would even want to invest in CSR / ESG style innovations? Of course, one obvious answer is that it is the right thing to do, from an ethical and moral perspective. Beyond such arguments, which we consider to be self-evident and implicit throughout this entire discussion, one might also point to the obvious public relations benefits that come from being known as an ethical or 'trustworthy' technology company. This is especially true when talking about AI technologies, which are associated in the popular imagination—rightly or wrongly, accurately or unfairly—with various doomsday scenarios, fueled by Hollywood films like the "Terminator" or Stanley Kubrick's Space Odyssey series, apocalyptic visions of mass unemployment and the reduction of humanity mindless executors of unfeeling AI overlords. AI companies have a tangible interest in being perceived as purveyors of a trustworthy brand of AI. This can translate to real market value, and investors – even those focused ruthlessly on profit and growth – must still recognize the importance of ESG as a key determinant of that hoped-for success in consumer markets. Finally, as some of our case studies show, ESG thinking can also open new market opportunities that may not have been obvious from a strictly growth & profit-focused orientation.

Applying the HRBA@Tech Model to AI Startups

The analysis that follows can be thought of as a filter. The full HRBA@Tech model is written to apply to all new and emerging technologies (NETs), across the entirety of their product lifecycles, and addressing the roles of all relevant stakeholders.

In this analysis, we are speaking only about the responsibilities of startup companies during the early phases of their lifecycle when they can still be accurately described as "startups." Furthermore, we are only speaking about AI-based technologies, which also modifies the product lifecycle somewhat, given the unique processes involved in bringing AI products to market. Thus, what was described at length in the 2022 Framework Paper can be described here with more brevity and greater specificity.

To effectively structure this analysis, we will approach it in the reverse order in which it is presented in the Framework Paper. We will start with "The Who" discussion (since we are dealing only with startups), then move on to "The How" discussion (to fixate only those parts of the project lifecycle likely to transpire while a corporation can still call itself a "startup"), and finally end briefly with "The What" discussion (reminding ourselves of the general principles that undergird the HRBA@Tech model). Using this as our baseline, we will then conclude with some research questions that arise from this application of the HRBA@Tech model to the AI startup ecosystem.

Stakeholder Analysis ("The Who")

Pages 98-99 of the HRBA@Tech report succinctly summarize what corporations, including tech startups seeking to bring AI products to market, should do to ensure that their products contribute to greater respect for human rights and social well-being. That text is reproduced and distilled for brevity and emphasis here:

The private sector is the primary driver of technological and scientific innovation today, and is therefore central to the development and deployment of NETs. This includes "tech-giants"—major corporations often with annual profits rivaling GDPs of mid-sized developed economies—but also other tech companies of various sizes including start-ups. [. . . .] Private actors typically exist to generate profits. The HRBA@Tech model in this report is built with this reality in mind, and attempts to balance these competing interests while ensuring that the development and deployment of new and emerging technologies is nonetheless better positioned to protect and promote human rights. While private actors do not have direct obligations under international human rights law, over the years there has been growing recognition of the crucial role they play in the advancement and realization of human rights and the need for corresponding responsibilities leading to efforts to accommodate private actors within the international human rights framework. In this regard, the [Guiding Principles on Business and Human Rights (UNGPs)] provide an authoritative framework for the corporate responsibility to respect human rights and a reference point for companies involved in the development and deployment of new and emerging technologies.

In light of the truly transformative potential that new and emerging technologies hold, [the 2022 Framework Paper describing the HRBA@Tech model proposed] that a "do no harm" approach is no longer sufficient. The HRBA@Tech model [. . . .] suggests moving beyond the UNGPs to embrace the possibility of actively crafting NETs to put them in service of human rights, or – to put it simply – to "make the world a better place." The HRBA@ Tech model recognizes the interests and constraints faced by private actors, and acknowledges that the promotion of human rights must always be weighed against the prerogative to continue generating profit. The HRBA@Tech model need not be antithetical to the interests of private actors and companies or incompatible with most existing business models. Private enterprises can incorporate various processes within the HRBA@Tech model directly into the [Technology Life Cycles (TLCs)] of NETs, either on their own or jointly with other stakeholders. At the outset, private actors must:

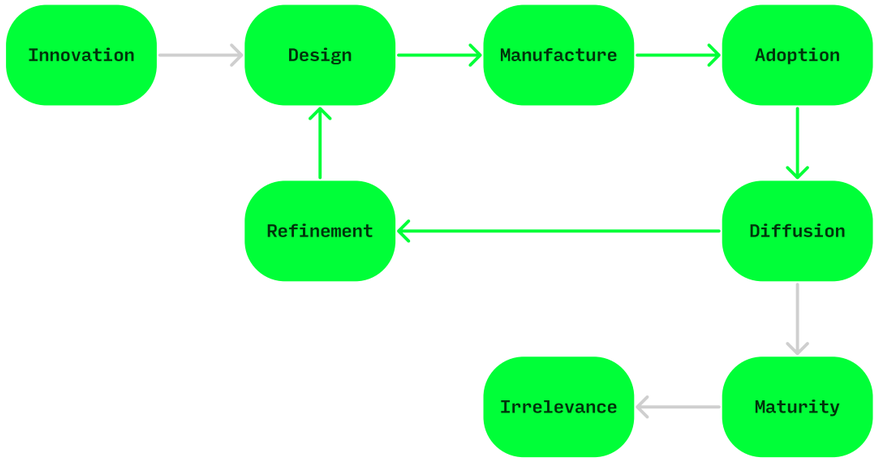
1. Comply with the standards articulated in the UNGPs.
2. Publish a formal company policy articulating its strategy for protecting and promoting human rights.
3. (for those enterprises intending to develop socially beneficial technologies) Clearly articulate the intended human rights objectives of a proposed NET and the company's strategy for achieving them. Entrepreneurs promoting such technologies should embrace a "human rights by design" process.
4. Conduct human rights due diligence and impact assessments to ensure that the new and emerging technologies do not, even inadvertently, harm people. These assessments should pay particular attention to constituencies that may be particularly vulnerable to the impacts of NETs, including those who may wish to opt out of its use.
5. Adopt a futures thinking mindset throughout the development and deployment of new and emerging technologies.
6. Ensure the safety of proposed NETs by ensuring the incorporation of safeguards or guardrails including "emergency brakes" during the design stage.
7. Be cognizant of relevant general or industry standards, including voluntary codes of conduct, and ensure that any NETs they promote have been designed and developed in adherence to such standards.
8. Ensure that human control over the technology remains meaningful, even while embracing the considerable upsides of such technologies.
9. Ensure that the technologies they develop are also used by others (subcontractors, clients, consumers, and licensees) in ways that are consistent with their intended use.
10. Make proactive efforts to be transparent about NETs (and their expected impacts on other stakeholders) throughout the various stages of the Technology Life Cycle (TLC).
11. Designate someone to answer questions and handle potential complaints regarding the development or deployment of an NET.
12. Put in place a grievance mechanism structured in line with the principles detailed in the UNGPs, in addition to other monitoring and oversight avenues.
13. Put in place mechanisms and protocols to suspend or alter the design of an NET should monitoring or user grievance analysis suggest that serious human rights impacts are occurring because of an NET.

The excerpted text summarizes in one place what private actors should do to ensure that their products are nudged in the direction of human rights and socially beneficial outcomes.

Product Lifecycle ("The How")

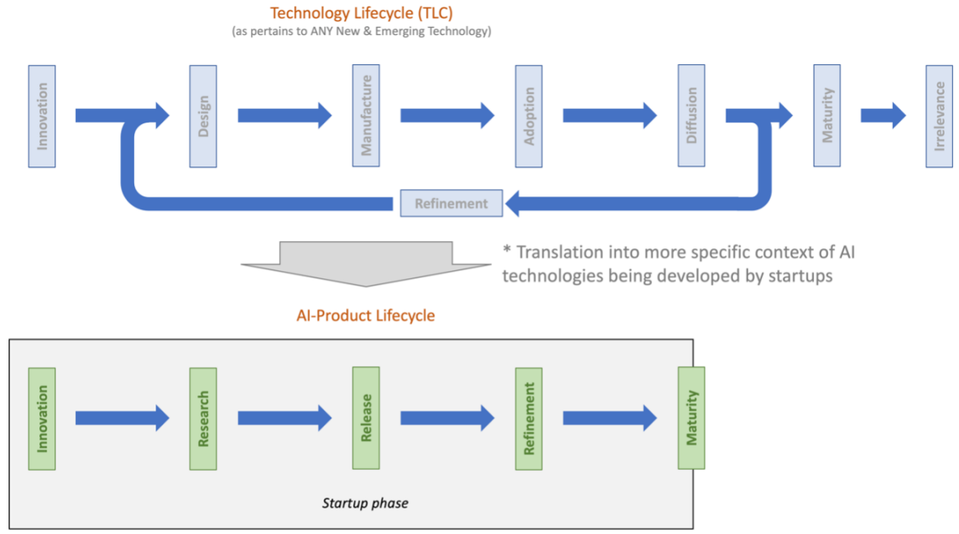
Turning to the next layer of this filtering process, it is clear that the products being designed by tech startups tend to be situated along the earlier phases of the Technology Life Cycle (TLC). To use the abstracted language of the Framework Paper, which proposed a TLC that could hold for all conceivable kinds of NETs, we are defining tech startups as those companies located at either the innovation, design, manufacture, adoption, or diffusion stages of the TLC. Once a company begins to refine its products or once their products can be said to have reached maturity, we are definitionally designating them as no longer a 'startup.' This may contrast with the way that may in those same companies might describe themselves. Google, Microsoft, Meta, Amazon, X and other "big tech" players might, for example, still encourage their employees or investors to think of them as places with a "startup culture."[149] This may be smart messaging to ensure a company remains nimble and innovative regardless of its size, but it does not bring them back into our definition of a "startup" for purposes of this paper. Rather, we might imagine these companies as finding themselves either at the Maturity or Refinement stages of the TLC, depending on whether they are actively seeking to re-invent themselves and position themselves into a new cycle of growth and innovation, the way Meta famously rebranded and reinvented itself away from just "Facebook" (a 'mature' social media company) to "Meta" a company focused on an NET (the "Metaverse").

Lifecycle for New & Emerging Technologies: From Innovation to Product Maturity



A first necessary step is to "translate" the above Technology Lifecycle Diagram into terms that make sense to those working with this specific technology of AI. In discussions with AI researchers, this following "translation" chart seems to make the most sense. It takes the broad stages and concepts articulated in the HRBA@Tech model and roughly equates them to how a tech company might describe itself as it evolves and grows from the initial innovation phase to maturity.

149. Mollar, Rany and Shirin Ghaffary, "Your Favorite Tech Giant Wants you to know it's a Startup Again," (Jan. 30, 2023), Vox, https://www.vox.com/recode/23573432/tech-layoffs-cuts-startup-facebook-google-amazon.



Changed are the terms we use for the Design Phase, which AI researchers might typically describe as the Research Phase of an AI-Product Lifecycle, where researchers test to see whether an emerging AI product functions the way it is intended.

Many AI products do not necessarily require a Manufacture Phase, since they are largely software based and thus can roll out immediately and at scale via the internet. There are exceptions, of course, especially when AI products are integrated into physical hardware products (such as in the field of robotics, or autonomous vehicles, or even AI products being incorporated directly into consumer electronics). In all those cases, the traditional Manufacture Phase postulated by the HRBA@Tech model would continue to exist.

For those AI products that do not require manufacture, a founder might move directly from the Research Phase towards the Release Phase, where the product is first released for customers or users to use. This might best be equated with the Manufacture Phase for NETs requiring a physical product.

The product would next go into a Refinement Phase, which typically takes place as users begin to use an AI product, and potential quirks or unintended consequences of an AI system begin to become more widely known. An example of this is the famous instance of a tech columnist somehow tricking OpenAI's chatbot into declaring its love for him and encouraging him to end his marriage to his wife.[150] Such instances allow an AI tech startup to make necessary adjustments to the AI system to make it less and less likely that bad actors can somehow break the guardrails on the system.

Eventually, a system will reach Maturity, at which point an AI system may be reaching a sufficient presence in the market that the startup can contemplate moving towards an initial public offering (IPO) or otherwise become established as a secure presence in the relevant target market.

While the need to continue focusing on human rights and the impacts of an AI system on potential stakeholders continues even after a company has 'graduated' from being a startup, our discussion of the HRBA@AI application will end once a product has reached maturity, since we are focusing only on the obligations incumbent to AI startups in this analysis.

What then, are the recommendations that attach to tech startups in the early phases of the TLC? Here again, we will summarize the 2022 Framework Paper, focusing on a sequenced discussion of when during their development trajectories AI Startups should be undertaking which efforts to ensure compliance with the HRBA@Tech model. The 2022 Framework Paper walks through these discussions in detail in Chapter 4, but this discussion filters that to apply only to AI startups.

150. Kevin Roose, "Bing's A.I. Chat: 'I Want to Be Alive.'" Feb. 16, 2023, New York Times, https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

Innovation

The Innovation Phase is both the most exciting but also the most daunting for a potential entrepreneur. Every technological innovation relies crucially on an initial moment of ingenuity, but it is also common knowledge that most good ideas never get off the ground. Founders during this phase of the product lifecycle must first come up with an idea or concept for a new use case for AI and next identify a clear market need for that idea. They must also conduct basic research to validate that their idea is at least scientifically conceivable and that their vision for a market case is at least plausibly viable. Building on that research, founders craft a business plan complete with their business strategy, goals, rough financial projections, and operational requirements to make their idea a reality. None of these plans would typically be fixed in stone, but they are necessary to get the attention of a mentor or accelerator program or VC investor who will then help to refine and tighten those initial ideas.

These activities typically occur in a rather casual and ad-hoc context, drawing on founders' personal skillsets and resources, and testing their ability to solve problems on their own. Some founders have business degrees or mentorship relationships to draw upon, but many do not, relying primarily on resources they can find themselves to drive their process forward.

During this Innovation Phase, the HRBA@Tech approach requires of founders to answer for themselves the question of whether their innovation will aim only to make money, or whether they might also aim to "make the world a better place" because of their innovation coming to market. If they decide on the former, they can largely restrict their thinking about human rights to the "do-no-harm" pillar of the HRBA@Tech model. If, on the other hand, they intend for their technology to also make the world a better place, the recommendation would be for them to spend time at the earliest possible innovation phase to describe in detail specifically how their technological innovation would do that. In that latter case, those specifics will form the core of their startup's business case and would be of great interest to potential investors. Having a robust vision for how their technology can improve the world around them will likely attract a whole new category of ESG-motivated investors who might not otherwise be interested in supporting a startup.

> "Our vision is to make AI more accessible to the general public. More specifically, our mission is to enhance the expressive abilities of today's youth, so that they can express their ideas more efficiently and make a real impact."
>
> **Seyoung Lee (CEO of Wrtn Technologies)**

A final note pertains to those tech startups promoting a product or service that could be used to deliberately inflict or cause harm to people. These might be, for example, engineers promoting AI solutions with military or defense applications, or AI products used for predictive policing, surveillance, or the prison industrial complex, among other examples. Such use cases are of course legal, and many technologists and entrepreneurs feed into these industries. The HRBA@Tech approach asks only that AI tech entrepreneurs promoting such use cases be certain that the end-users of their technology do not intend openly to use it in ways that harm or undermine human rights.

HRBA@AI Intervention Vectors during the Innovation Phase			Resources (time & effort) required for an AI startup to do this properly
Do No Harm	Risk Management Brainstorming Session	What kinds of general risks might a particular AI use case entail, and are there any obvious ways even at the earliest innovation phase that one can think of to eliminate, reduce, or mitigate those risk in some way?	Low
Make the World a Better Place	Incorporation of Social Mandate into Business Plan	If the intention is specifically for an AI-enabled product of service to make the world a better place, an entrepreneur, at the earliest phase possible, ought to be able to convincingly show how specifically it will do that. Investors and mentors will want to see that vision clearly presented to them before they commit. But when included, this vision might also serve to attract a new class of values-oriented investors, including those concerned about their ESG investment portfolio.	High
Potentially Harming People	Due Diligence	If a technology is one that obviously has the potential to harm or undermine people's ability to enjoy their human rights, an investor or tech startup ought to be sure that the government who will ultimately take charge of that technology is one that respects human rights and does not obviously intend to use a new technology to undermine human rights. Investors should do this as much for ethical reasons as also to protect themselves from legal liability should they be accused of conspiring to promote the agenda of a human rights violating government.	Medium

Depending on a tech entrepreneur's assessment of the ultimate potential of their AI-enabled product or service to either promote or harm human rights, they ought to engage in different thought processes at this initial Innovation Phase of the product's development.

Do No Harm

Brushing aside the florid rhetoric that often is associated with most corporate mission statements, many – if not most – startup founders are ultimately motivated by the desire to generate wealth. First it must be said that there is nothing illegal or nefarious about such a motive, especially in a capitalist system where innovation, wealth creation and ultimately social progress is thought to be driven by this profit motive. Even profit-driven enterprises, however, still need to fundamentally respect human rights. The social contract permits them to pursue private profits so long as they do not harm the legitimate rights of others in that process.

If founders are pursuing an innovation that they believe will generate profits, and if they consider that profit motive to be the primary justification for the innovation, the HRBA@Tech model requires of them only to engage various "Do No Harm" efforts (labeled in blue below).

Voluntary Risk Management Brainstorming Session: During the innovation phase, a commitment to "Do No Harm" requires that the founders engage in a "voluntary risk management brainstorming sessions as part of the investment cycle" with the identification of cost-effective "risk mitigation" strategies as the sole objective for those conversations (HRBA@Tech, 2022:80). This may sound like quite little to ask of tech startups at this early phase of their trajectory, but the claim is that it is sufficient, at this very early stage, to set the switch of a startup – even an exclusively profit-focused startup – towards a culture of respecting human rights and social responsibility. This simple brainstorming session can implant the seeds of the HRBA@Tech approach directly into the DNA of a fledgling corporation, with very limited time or budgetary requirements for the founders.

> "After we applied filters to our chatbot services to weed out 'harmful' content, the number of users decreased. However, we had to make conscious efforts to refine our AI model when we learned that our service could be used by bad actors to cyberbully or circulate pedophilia-related content."
>
> **Junghoi Choi (CEO of Simsimi)**

Make the World a Better Place

Many founders do indeed hope for their products to make the world a better place, but many also leave those assumptions and aspirations unstated. The pressures of finding funding for a new product tend to relegate the specifics of how an NET will make the world a better place to an afterthought at best, or perhaps something best left for the public relations team of a new product.

Neglecting the early opportunity to articulate these specifics, however, also leaves possible values-driven sources of investment untapped. For any AI entrepreneurs wishing to have their products contribute to an improvement in human rights, therefore, this earliest Innovation Phase is the ideal moment to define that pro-social business plan with clarity.

Incorporation of Social Mission Directly into the Startups Business Plan: If the founders intend for their AI-enabled product or use case to make the world a better place, the HRBA@Tech model requires them to prepare for a "careful vetting by investors and/or corporate or government sponsors of proposed business plans to ensure that they clearly articulate how the NET will work to 'make the world a better place.'" (HRBA@Tech, 2022:80) Founders purporting to introduce technologies they claim will make the world a better place should be ready to anchor those claims not just in aspirational rhetoric, but also in the products' tangible business models. If the social benefits of a technology are central to that product's identity, the founders ought to be prepared to make their pitch about how those social benefit will flow from the innovation.

Potentially Harming People

The HRBA@Tech model does not preclude the reality that founders may legitimately develop technology destined for use by law enforcement or military users, nor does it stigmatize such innovation in cases where founders have no indication that the end-users of those innovations would abuse the technology or use it to intentionally violate human rights. An AI-enabled system designed to identify dangerous terrorists, for example, might be appropriate if used by a government known for its meticulous respect for human rights. But that same technology deployed by a country known to label political opponents as terrorists, for example, would be inherently suspect. Founders owe it to themselves to reassure themselves, ethically and legally (see Lungisile Ntsbeza et al v. Daimler AG et al., 2009) to not be willfully complicit with state-sponsored human rights violations.

Due Diligence: If the AI use case is designed specifically to prevent people from self-effectuating (for example, by limiting their freedom, violating their privacy, or even killing or injuring them), for example in the case of AI technologies with military applications, the HRBA@Tech model requires the founders, as a matter of personal responsibility, conduct "due diligence of [the] government sponsors [or potential users of the technology]."

Research

The 2022 Framework Paper noted that "[t]he design phase is one of the most important intervention points for injecting human rights considerations into an NET. At this point, after the kernel of a new technology has been developed but before the final form that a new technology will take has been ossified by virtue of a manufacturing process, human rights considerations can still be mainstreamed with relative ease into a new technology" (HRBA@Tech, 2022:80).

> "As start-ups are always suffering from a lack of resources and face the pressures of survival, some may be skeptical about startups' ability to raise and solve ethical concerns, saying that AI ethics is only an afterthought for them. However, AI start-ups could also be well positioned to realize a higher level of AI ethics standards than the more established players in the market who may be complacent in complying with the existing legal regime. Whereas the bigger tech companies would prefer that the existing rules stay unchanged, it may be the destiny for startups, and perhaps in their best interest, to break away from existing, sub par practices, thereby leading the way forward."
>
> **Gene LEE (CEO of LBox)**

During this phase, startups prepare to launch their business by securing an initial round of seed capital – perhaps from a VC investor or by joining an accelerator program of some sort. This might be when the business incorporates and builds a core team of founders, a board of directors, etc. It might also be when the business begins to develop a viable product – something it can use to gather feedback and generate interest among investors and potential customers about the product or service.

The 2022 Framework Paper suggests that AI Tech startups at this phase can do several things to ensure that their final products or services remain consistent with human rights priorities:

HRBA@AI Intervention Vectors during the Research Phase			Resources (time & effort) required for an AI startup to do this properly
Do No Harm	Futures Thinking + Vulnerability assessment	How will a technology – when it reaches full maturity and adoption by all relevant audiences – impact the most vulnerable in society, and can those vulnerabilities be somehow lessened by means of concrete design innovations in the technology itself before it goes to manufacture?	Low
	Security Research	Careful stress-testing of the technology to ensure that there are no unintended consequences – even inadvertent or accidental ones – that could jeopardize the rights or well-being of impacted stakeholders.	High
	Grievance Processes	Design of grievance processes that can accompany the technology and that can serve as 'hazard indicators' in cases when the technology does not meet its stated purposes or when the technology inadvertently leads to negative consequences for rights-holders.	High
Make the World a Better Place	Technology Transfer	Technologists should already be thinking during the design phase how their technologies will be shared with underserved markets as part of the technology diffusion phase. This may be part of the company's future ESG strategy but should be planned by design.	Low
	Transparency	If the design team hopes to use the technology to actively empower vulnerable populations, it would need to first 'translate' the projected impact of that technology into terms that a vulnerable population can understand.	Medium
	Due Diligence	Designers of NET should consult with vulnerable communities and any other potentially impacted groups to not only ensure that their technology 'does no harm' but that it also 'makes the world a better place.'	High

Some of the above processes require a significant investment of time and resources. Not investing in those processes, however, leaves a fledgling startup vulnerable to potentially existential consequences down the road, which should also be taken by investors and

early adopters as a sign of unacceptable risk during this early phase of a startup's development trajectory.

Some of these activities fall under the "Do No Harm" heading and should therefore be done no matter what the startup's business model is. According to the HRBA@Tech model, these activities would be required to minimize the risk that a startup's operations, products, or services inadvertently harm the human rights of its employees, consumers, or host communities. Others fall under the "Make the World a Better Place" rubric and apply only to those startups that see it as part of their mission to improve the general wellbeing of those who interact in some way with its operations.

Do No Harm

Conduct a Vulnerability Assessment: Conducting a vulnerability assessment should be relatively straightforward. This is essentially a brainstorming exercise focused on the identification of potentially vulnerable communities or individuals who might conceivably be harmed by a startup's products or services. Startups should always be thinking about how its products impact potentially vulnerable communities as part of their commitment to the "Do No Harm" principle. What to do with this assessment is then left up to the managers of the startup. Some may wish to actively consult with those vulnerable communities to co-design the products or services. Such consultation would move away from risk management and towards a commitment to 'make the world a better place.' (see below).

> "We take privacy seriously. We have rigorous discussions on how to de-identify personal information so that the rights of data subjects are not unduly infringed upon. It is an important task, but an arduous and costly one as well. In this regard, if governments could find a way to disseminate sound data sets to work on or provide practical guidelines, it will immensely boost the productivity of many AI startups in their research phase."
>
> **Jungkeun LIM (CEO of BHSN)**

Conduct Security Research and Testing: Security research is labor intensive. One tech startup we spoke to devoted approximately half of its staff during the Research Phase toward developing guardrails to train the AI-enabled system to reliably achieve its outcomes. This investment, however, is essential for startups seeking to earn the trust of skeptical consumers. AI-enabled products still cause trepidation for many potential customers. If an AI startup wishes to achieve sustainable success, there should be a focus on safety early on. One negative story in the press about an AI start-up's products or services can often be enough to devastate that company's chances of success. Especially for startups, their ultimate success often hinges on their ability to make the case to customers that they offer a higher degree of security than currently available products on the market. From a business perspective and an ethical perspective, investments in safety research are crucial for early-stage tech startups in the AI sector.

> "AI is different form human intelligence in that the impact of its training is perpetually cumulative. An AI system trained on tainted input data produces tainted output data which again feeds back into the system. This is why I cannot stress enough the importance of establishing safe guardrails in the early stages, rather than later when the system begins to spiral out of control."
>
> **Byungjoon KIM (CEO of Hantech)**

Develop a Grievance Process: It takes time and effort for a company to design a viable grievance process, especially if that design process is open to input from relevant stakeholders. The Ruggie Principles contained in the UNGPs provide a clear vision for any such design process, articulating the principles around which any grievance process should be built. Numerous resources, including a thriving specialized consulting industry, exist to help corporations design grievance processes that are relatively simple to maintain but that also present significant added value to a corporation. Designing a grievance procedure that will grow with and contribute to a growing business operation is a wise investment. Developing such a grievance process is not a concession to pessimism (or "wokeness") or a sign that corporate management expects there to be a great number of problems to be resolved. Rather, it is a recognition of the inevitability, in any operation, that grievances or disputes are likely to arise at some point, and a recognition that when such grievances arise the organization ought to be prepared to handle them constructively rather than letting them spiral out of control.

> "As a C2C company, it is important for Daangn to immediately respond to customer complaints. Harnessing machine learning capabilities, our operations team detects hazardous content and develops tools for filtering content and determining the order in which content is exposed to users."
>
> **Mina JUNG (GR Lead of Daangn)**

> "An example of a tool currently under review is the "Feed Soundness Indicator(s)." The indicators would be used to classify content based on data used by customers or user feedback, quantify "ideal connections", and use them for filtering and recommendation in the future. For example, Daangn prohibits the sale and purchase of fish, and such decision is based on an analysis of the customer's usage data, where they found that the sale and purchase of fish usually takes place with the intent of mass trading for pets, a practice that frequently involves unethical business practices such as animal abuse."
>
> **Jaeyoon CHUN (Machine learning engineer at Daangn)**

Grievance processes are useful for human as well as technical concerns. They can surface the kinds of issues that can plague any operation: issues of harassment, discontentment among the staff, sexism, racism, ageism, etc. But grievance processes can also serve a technical function, in that they can provide managers with early warning if the goods and services of a company are causing unintentional harm among its users. Not having such an early warning system risks having those issues play out in much more costly ways through the media or litigation. In such situations it is always better to have a robust grievance process in place, waiting to be used, rather than to be caught unprepared when aggrieved stakeholders find that there is no grievance process available to resolve their concerns.

Make the World a Better Place

Anticipate and Plan for Future Technology Transfer: A key component of making the world a better place is a commitment to equity, both within and across nations. The globally relevant corpus of human rights includes a commitment to the global sharing of knowledge and scientific achievements. When these achievements are made by corporations, and shielded by robust intellectual property regimes, such knowledge sharing typically only happens in line with a corporation's profit function. If a particular society or customer is simply too poor to afford a certain AI-enhanced product or service, the market dictates that they will simply never have access to that good. A commitment to making the world a better place, therefore, requires devoting some thought towards a corporate strategy that will eventually open opportunities for communities to benefit from technological innovation even if they lack the resources to compete on the open market. Anticipating such technology transfer does not necessarily require a non-for-profit business model, but it does require some strategic planning about how to nonetheless protect legitimate trade secrets and intellectual property protections.

At this early Research Phase, a focus on technology transfer requires little more than a brainstorming process and visioning exercise. The results of that process can then be incorporated into the business plan for the AI-enabled product or service. Companies that think about eventual technology transfer can claim to be planning a tangible contribution towards 'making the world a better place.' This brainstorming activity is best done early on in a startup's development trajectory, as it will help it attract certain types of ESG funding and implant a medium-to-long-term strategy for sharing of new technologies with underserved markets directly into the startup's business plan.

Transparency: Transparency is frequently mentioned as an integral part of AI ethics, applicable to all AI use cas-

> "Some critics worry about the displacement of authors. More specifically, they raise concerns that the prevalence of AI-assisted writing will diminish the value of expressiveness in works of authorship. However, we believe that an author's unique expression, or an expressive element in a literary work will remain, albeit in the territory of luxury goods. Our AI service will be liberating in the sense that most people will get a chance to focus more on the raw ideas and imagination while being assisted by AI as to how they can best express them."
>
> **Seyoung LEE (CEO of Wrtn Technologies)**

es, not just those seeking to make the world a better place. For AI tech companies pledging themselves to 'make the world a better place,' however, transparency requires not just transparency in response to applicable laws, regulations, and court orders ("responsive transparency"), but rather proactive transparency. The 2022 Framework Paper distinguished these two levels of commitment to transparency. All AI startups, regardless of their commitments to ESG principles, should commit themselves to responsive transparency as part of their commitment to simply follow the applicable laws.

Proactive transparency, on the other hand, requires businesses to think about the barriers that might exist between company insiders and the communities that interact with their products and services in some way. Cultivating this kind of transparency takes some time and effort, but ultimately also has payoffs down the line in terms of stakeholders' understanding of a start-up's products and services.

> "We use content moderation as an opportunity to spur discussion within the community. We ask our users for their opinion and reflect their feedback in our operational policy. However, often the users seem split on the issue, and line-drawing becomes very difficult."
>
> **Jay LEE (Founder and COO of Triplecomma) & John HAN (CTO of Triplecomma)**

Consulting with Vulnerable Communities: Consulting with communities requires a significant investment of time and resources. An early investment during the Research Phase of the product lifecycle can yield major payoffs down the road. A proactive effort to consult with obviously identifiable vulnerable communities can prevent accusations later that a company was acting without regard for the impact of their decisions on local communities. More importantly, consultation can give communities a sense of agency and co-ownership over a product that can also positively impact the startups' later chances of capturing market share. Here too, an early investment can yield major future payoffs, and should also serve as a demonstrable signpost to investors of a company's stance on ESG principles.

Release and/or Manufacture

The "Manufacture" stage uses a terminology best suited to the development of physical technological devices and presumes a process of actually making those products and bringing them to market. SaaS AI use cases do not require the physical "manufacture" of any products, but they do require substantial programming and refinement of various software-based products or services. AI technologists might describe this as the Release Phase, when a product is first released to customers and consumers. This stage usually involves a significant marketing and promotional push to attract customers and establish an early market presence. As startups prepare to launch, accelerators and VCs can provide valuable support by offering access to networks, industry partners, and investor communities. They can also assist with early-stage marketing and sales strategy, helping to create buzz and momentum as the product launches.

The 2022 Framework Paper suggests that AI Tech startups at this phase can initiate several processes to ensure that their final products or services remain consistent with human rights priorities:

HRBA@AI Intervention Vectors during the Release and/or Manufacture Phase(s)			Resources (time & effort) required for an AI startup to do this properly
Do No Harm	Environmental, Social, and Governance Policies	ESG policies and procedures should be built into the manufacturing phase of any NET.	Medium
	Supply Chain Management Policies	Responsible and ethical supply chain management policies need to be enacted to ensure that there are no negative human rights impacts at any stage of a NET's supply chain.	Medium
	Fair Labor Practices	Companies developing NETs need to ensure that they have fair labor policies and that they are not violating the human rights of their employees.	Medium
	Establishment of Internal Grievance Processes	Internal grievance process for employees are necessary for any company working on an NET, so that its employees can file internal claims for issues such as discrimination or unlawful practices.	Low/ Medium
	Independent Monitoring	There should be independent monitoring of the manufacturing stage of all NETs, and in some cases both internally and externally.	Low/ Medium
Make the World a Better Place	Civil Society Involvement	Civil society should be consulted and involved in the manufacturing stage, whether as independent observers or other types of participants.	Medium

Most of these processes fall under the "Do No Harm" pillar, and are therefore incumbent on any startup, regardless of whether they have set for themselves the goal of "Making the World a Better Place." Indeed, each of these five processes, broadly speaking, fall under the heading of ESG Policies, and increasingly should be considered as standard elements of doing business. These processes can be labor intensive, but should also be considered normal and necessary aspects of a startup growing and maturing to become a viable market presence.

Do No Harm

Environmental, Social & Governance Policies: Implement-ing robust ESG policies in a company involves several key steps. The first is to understand ESG as a concept and its importance, and more specifically how the startups business model might interrelate with those con-

cepts. AI companies might, for example, think about the environmental impact of their servers, and seek ways to minimize the carbon footprint of those operations.[151] Or they might think about social issues such as their own internal labor practices or the impact they might be having on certain communities in which they operate. Finally, they might think about governance issues having to do with corporate ethics and transparency. These initial surveys are best done with the help of outside stakeholders, which is one area where corporations might choose to involve civil society (see below, under "Make the World a Better Place"). While still unusual in the business world, bringing outside perspectives into these discussions can help give a much more realistic sense of a business' potential impacts on communities, and therefore offer valuable insights into how a business might minimize those negative impacts.

After conducting a baseline assessment, a firm might want to set for itself some clear and realistic ESG-related goals that it would like to achieve for itself. These goals should not be seen as charity, or add-ons living in awkward tension with an existing business model. Rather, they should be adjustments to improve environmental impact, improve a company's social impact, and improve its governance structure that would be aligned with the company's overall business model and perhaps even enhance it. Depending on the size of a company, management should designate a specific unit, department, or compliance officer as the champion of human rights within a company. This unit should have the technical expertise to engage in constructive problem solving with relevant other units and focus on promoting a cross-cutting regard for human rights as a central part of the corporate culture.

> "Having a governance structure is crucial because, in the real world, people require designated authority figures. Assigning names to teams and personnel, along with providing titles, is necessary for individuals to fulfill responsibilities and be motivated to carry out their duties. This significantly contributes to creating value and influences how effectively you can advocate for AI ethics issues when dealing with other departments."
>
> **Jinhwa HA (Manager of Kakao Human Rights and Tech Ethics Team)**

After having set specific and achievable goals, a company should then formulate policies to begin mainstreaming its ESG commitments throughout its operational model. A few areas are particularly important to think about for most companies:

Supply Chain Management Policies: To the extent a company produces physical products, it should develop strategies to ensure the ESG alignment of its suppliers to those it also wishes to uphold. As many garment, footwear and computer hardware companies learned the hard way, lax labor standards in supplier firms delivering only components of a final finished product can easily tarnish the reputation of the brand-name hardware producing company.

Fair Labor Practices: Somewhat related to supply chain management processes, but directed internally at one's own staff and employees, a robust commitment to ESG dictates a commitment to living wages and fair market practices. While this may not sound like a major concern in AI companies with its highly-paid technical experts and engineers, it does have implications for non-tech staff who work at the companies as well as the widespread practice within the tech industry of outsourcing certain high-intensity business operations to lower-cost markets (often in the global south) where workers earn substantially lower salaries and enjoy few of the benefits enjoyed by tech workers in headquarters.

Establishment of Internal Grievance Processes: The importance of establishing a robust grievance process was described above. In particular, during the "Manufacture" or "Release" Phase of an AI-enabled product or service, grievance processes can serve a crucial function of providing an early warning system if the AI, for whatever reason, begins to behave in unexpected or counterproductive ways. Firms should also establish internal grievance process for employees so that they can file internal claims for issues such as discrimination or other unlawful practices. As the business begins to grow, and as new stakeholders become involved, these grievance processes should be re-examined and potentially amended to ensure that those new stakeholders (for example customers who engage with an AI product in the marketplace after its public launch) also have viable access to a grievance process, that their experience interacting with that process still reflects the Ruggie Principles contained in the UNGPs, and that the company benefits from the early-warning role that such a grievance process should play.

Independent Monitoring: The company will also need to monitor its progress on these various ESG priorities. This means first establishing objective metrics of success, tracking those metrics, reporting them publicly, and periodically revisiting them to assess whether the metrics gathered truly reflect the spirit of the ESG goals the company set for itself. To support its monitoring mission, the company can and should invite external and independent auditors into the process to ensure the legitimacy of this auditing process.

> "Our monitoring mission starts from within. We emphasize the education and training of our staff from top to bottom and the monitoring of our system to ensure that our cybersecurity measures are always intact. Essentially, we have made deliberate efforts to make this kind of vigilance part of our culture. In fact, when it comes to security issues "basic hygiene" can prevent 90% of cybersecurity risks."
>
> **Chan YOON (Director, Corporate, External & Legal Affairs at Microsoft Korea)**

After these various ESG policies are developed, they need to be integrated and mainstreamed into a startup's overall business operations. This will require additional rounds of training, capacity building, and oversight. Here also external consultants, in-house specialists, industry consortiums, or partnerships with specialized civil society groups can be quite helpful. Especially those partnerships involving non-company insiders as facilitators can serve as clear and positive signifiers of transparency by the startup, as well as messaging to internal and external audiences that the company takes these priorities seriously.

Finally, the company will need to communicate about its progress vis-à-vis its ESG goals internally and with investors, but also – if it chooses to be proactively transparent –publicly with stakeholders in the community.

Make the World a Better Place

Civil Society Involvement: Many businesses are extremely hesitant to invite civil society into their operations to support their efforts to implement more realistic ESG commitments. These concerns are understandable. That said, for certain startups such partnerships can prove to be extremely valuable. They offer the opportunity for more transparent communication about issues that may be affecting workers or other stakeholders, articulated by civil society groups who are less concerned about the repercussions of speaking up to management. In exchange for this transparency, corporations are also able to claim publicly that they work in partnership with those civil society organizations to craft their ESG policies, and that the integrity of those organizations attaches also to the corporation's efforts to approach this process in good faith.

151. Melissa Heikkilä, "We're getting a better idea of AI's true carbon footprint," (Nov. 14, 2022), MIT Technology Review, https://www.technologyreview.com/2022/11/14/1063192/were-getting-a-better-idea-of-ais-true-carbon-footprint/.

Refinement

For AI products, the adoption phase (which AI technologists might refer to as the "Refinement Phase"), is when a product has been launched onto the market and ideally a startup's products or services begin to "go viral." Success begins to beget success, and a product's reach begins to grow exponentially.

HRBA@AI Intervention Vectors during the Research Phase			Resources (time & effort) required for an AI startup to do this properly
Do No Harm	Due Diligence	Developers of NETs should conduct ongoing due diligence to protect from (and correct) negative human rights of their product(s).	Low
	Adjustments to Prevent Distortion	The necessary adjustments need to be made by those creating and marketing technologies to prevent distortion.	Low
Make the World a Better Place	Ethical Marketing	Ethical marketing requires a focus on not only how the NET benefits customers, but also how it 'makes the world a better place,' by, for example, benefiting socially or environmentally responsible causes. It includes avoiding false or misleading claims or representations of the product.	Low
	Outreach & Maintenance of Grievance Procedures	Companies should conduct outreach to potentially affected communities and groups and maintain grievance procedures for anyone negatively affected by their products.	Low

If a startup has undertaken the higher-intensity efforts during the Research and Release / Manufacture Phases of the product lifecycle, relatively little additional work effort will be required of management at this stage. Most of the processes listed above will be largely passive, and the institutional and cultural infrastructure necessary for these processes to operate effectively will have already been put in place earlier.

Do No Harm

Due Diligence: At this point, an existing grievance process, and existing investments in the ESG priorities and policies of the company should be sufficient to allow management to monitor, passively but in an ongoing manner, whether any new human rights challenges have arisen because of the company's rapid expansion.

Adjustments to Prevent Distortion: When the company identifies potential issues, it would then need to expend efforts to course-correct. Given previous investments in a corporate ESG philosophy, however, these course corrections are likely to be rather modest.

Make the World a Better Place

Ethical Marketing: Ethical marketing practices, especially in the context of products hard-wired to make the world a better place, are essentially synonymous with regular marketing practices. At this point, the corporation's efforts to promote human rights in fact feed directly into the company's approach to advertising.

Outreach & Maintenance of Grievance Procedures: Finally, a company would need to engage in ongoing maintenance of its grievance procedure – again promoting it and ensuring that it follows the company's inroads into new markets. This way, the grievance system can continue to serve as the company's primary early-warning system in case its operations are causing unintentional harm.

Deliberate efforts by startups to cater to user needs can ultimately pay dividends. Although efforts to infuse human rights considerations into their system may be burdensome, a number of startups in their Refinement Phase seem to have embraced the cause as a springboard to success and sustained growth.

> "I think the most important objective of startups is "user satisfaction." If there is a company conflicted between 1) increasing advertisement sales exposing feeds based on metrics that attract users with instant gratification, and 2) adopting soundness indicators at the expense of such sales, I would suggest thinking about what you believe is the proper direction of a 'sustainable platform' from the perspective of your customer base. You should prioritize ensuring sustainable user satisfaction, and then think about how to better monetize based on this during the decision-making process."
>
> **Woochul PARK (Agenda Research Leader at NAVER)**

> "Excessive control over content could make the platform less exciting, but on the other hand, blind neglect toward collective pleasure-seeking behavior is likely to deteriorate the quality of user experience. Some vocal users tend to advocate free speech and defy efforts to moderate content. In terms of user retention in the immediate term, we would be better off freely exposing more exciting content, but we have also witnessed the rise in user fatigue due to excessively stimulating content.
>
> "Striking the right balance with content moderation is also closely aligned with our profit motives. Avoiding content moderation may have a positive impact on short-term sales, but it may have a lingering negative impact on our revenue and reputation in the long run."
>
> **Jay LEE (Founder and COO of Triplecomma) & John HAN (CTO of Triplecomma)**

Maturity

The final phase of the product lifecycle that can arguably still pertain to startup enterprises is the diffusion stage (which an AI technologist might refer to as the "Maturity Phase"), where a startup's product begins to be incorporated into other products produced by other corporations. One startup's product might, for example, have become so successful that other applications incorporate it as the foundation for their own subsidiary products (perhaps paying the original startup royalties to do so). In such situations, the company relinquishes some control over its original goods or services and begins instead to collect royalties for its intellectual property from other service providers.

HRBA@AI Intervention Vectors during the Maturity Phase			Resources (time & effort) required for an AI startup to do this properly
Do No Harm	Vetting of Potential Licensees	Potential licensees of NETs must be thoroughly vetted to ensure that they will not use the licensed technology to harm others or commit human rights violations. While an 'owner' of any particular NET cannot ensure with complete certainty that a licensee will not use the NET in a harmful manner, licensors should engage in practices such as reviewing the human rights records of potential licensees (e.g., in the example of a government licensee, examine the government's human rights record) and the licensee's stated desired use of the NET.	Medium
Make the World a Better Place	Due Diligence	After licensing or otherwise diffusing an NET, the owner (as well as the licensee) should conduct ongoing due diligence to ensure that the technology is not having any potential negative human rights impacts.	Low
	Installation of Technological Safeguards to Prevent Abuse	As detailed in Chapter 3, guardrails should be installed within any NET to prevent negative human rights impacts.	Low

The corporation may relinquish primary responsibility for the AI product at this point, as it hands over control and use of the technology to another service provider. Depending on the circumstances, there may be little left to do for the startup at this point to remain in line with the HRBA@Tech approach.

Do No Harm

Vetting of Potential Licensees: To "Do No Harm," a corporation must merely assure itself that the subcontractor or licensee of a technology not have an obviously human rights violating purpose in mind for obtaining a license for the technology. This requires some modest vetting of their business model and track record as a business or public authority so far.

In the vetting process, however, it is advisable for startups to be at least open-minded about the purpose that their AI product may be used for. In other words, they should be careful not to be overly judgmental about the licensee's intended purpose for usage.

> "Content moderation is a delicate matter. By failing to capture the subtle nuances of our expressions, we can easily over-regulate. As we analyzed the use patterns of heavy users of Simsimi, we encountered a quite interesting case where a user would swear at the Simsimi chatbot for an hour nonstop. At first, we suspected this to be another abuse case, but on second thought, we came to realize that this unhappy individual may have been engaging in ventilation, which is a common treatment applied to psychiatric patients.
>
> "Should even the lightest of swear words be filtered out and uniformly regulated? We may have to pause for a moment and think about the effect of content moderation on the freedom of speech and the power of language as a ventilation tool."
>
> **Junghoi CHOI (CEO of Simsimi)**

The potential of chatbot-assisted health support is well-documented through various research conducted using Simsimi's anonymized interaction data.[152] Users sought health-related information and shared emotional messages with the chatbot, indicating the potential use of chatbots to provide accurate health information and emotional support.[153] Especially for users who have difficulty communicating emotions to other humans, these chatbots would provide helpful information about depressive moods and how to cope with them.[154]

On the other hand, the risks and dangers associated with these AI companions have also been reported.[155] The chatbots also had the potential to provide ill-informed guidance, magnify negative emotions, or inadvertently motivate harmful acts such as self-harm or harming others.[156]

Setting boundaries on what ideas we can tolerate and sympathize with, and how far we will allow them to be expressed in our interactions with AI would be important decisions we have to make down the road. Pushing the boundaries of AI ethics to ensure human dignity is upheld in every respect could lead to a far deeper level of sophistication of AI.

Make the World a Better Place

Due Diligence: For companies intending to make the world a better place, they can also continue to extend its monitoring and due diligence systems even to extend to the operations of the licensee. These due diligence processes would presumably have to be tied to the underlying AI use case or product and negotiated as part of an overall licensing agreement with the licensee.

Installation of Technological Safeguards to Prevent Abuse by Licensees: The final strategy a company can use before licensing its AI enabled products or services is to implement specific use guardrails into a product that guarantee that a product will not be used in a way inconsistent with its intended use. This is a technical solution, and it presumably limits the market value of a certain good or product, but it would go some way towards ensuring that even licensees stay bound to the "Make the World a Better Place" vision for a certain AI enabled good or service.

> "In building community standards, at least initially, we may have to be satisfied with abstract principles. As the standards evolve, the time will come when hyper-personalized value propositions or deeper nuances of ethics become important. At that point, more concrete and intensified discussions on training, filtering, and retraining would need to be made."
>
> **Seyoung LEE (CEO of Wrtn Technologies)**

152. Chin H et al., "The Potential of Chatbots for Emotional Support and Promoting Mental Well-Being in Different Cultures: Mixed Methods Study", J Med Internet Res 2023; 25:e51712, https://www.jmir.org/2023/1/e51712; Chin H et al. User-Chatbot Conversations During the COVID-19 Pandemic: Study Based on Topic Modeling and Sentiment Analysis, J Med Internet Res 2023; 25:e40922, https://www.jmir.org/2023/1/e40922.

153. Id.

154. Id.

155. Julian De Freitas et al., "Chatbots and Mental Health: Insights into the Safety of Generative AI,"(Oct. 26, 2023), Harvard Business School Marketing Unit Working Paper No. 23-011, at 5.

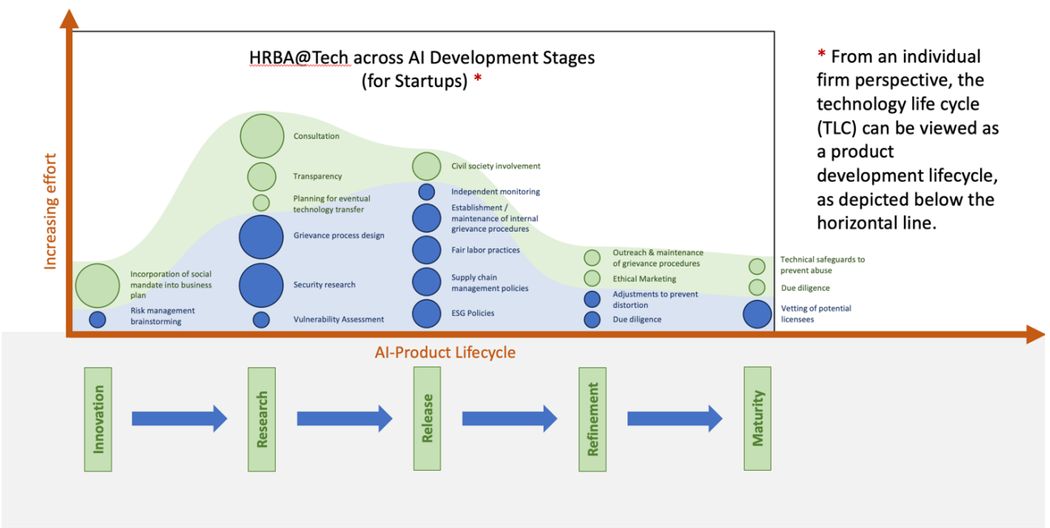
156. Id.

From Principles to Processes ("The What")

The analysis above already described in detail what startups should do, at what point in their product lifecycle, to be consistent with the HRBA@Tech approach. The principles that should guide those processes are implicit in what was written above. For companies that subscribe only to the "Do No Harm" approach – in other words companies who pledge to abide by the provisions of the UNGPs and standard business practices regarding ESG programming – they should be guided by the principles of Legality, Non-Discrimination, Safety, and Accountability & Access to Remedies. Any business developing AI-enabled products or services should embrace these principles as a 'bare minimum' standard to guide its operations. Businesses claiming to also make the world a better place, on the other hand, should also aspire to promote a commitment to the principles of Empowerment, Proactive Transparency, and Participation in their efforts to advance the cause of human rights.

Question(s) Presented

This chapter has featured a straightforward application of the HRBA@Tech framework as applied to an AI-product lifecycle. Taking the relative effort required to carry out each of these processes, one can chart the effort required by an AI startup over time, as the company evolves from the innovation phase through towards the Maturity Phase. This analysis shows that the effort required by firms pursuing a "Do No Harm" strategy of human rights risk mitigation is highest during the Research and Release Phases of the AI product lifecycle. The analysis further suggests that firms intending to "Make the World a Better Place" as part of their core business strategy must additionally expend substantial effort during the Innovation and Research Phases of the product lifecycle. While significant, we have found that these early investments pay off, especially as the corporations mature. Early human-rights based investments serve to prevent human rights-related crises from damaging the company's reputation and bottom line during later stages of the firm's AI-product lifecycle, when they would be far more expensive to remedy. For firms claiming to incorporate human rights objective directly into their business model, early investments are crucial to attract the support the startup will need to get off the ground and find the necessary user base that would help the company flourish.



Our case studies explored the extent to which these theoretical ideas were reflected in the actual practice of AI startups, and whether there might also be additional processes not foreseen in the HRBA@Tech model that corporations used to protect and promote human rights.

Analysis: Integrated Lessons Learned from the Case Studies

Ten lessons stand out from the above analysis:

1. **The term "tech startup" is too broad to serve as a useful category when discussing what a tech startup should do at which point of an AI product's lifecycle.**

   The term "tech startup" is almost absurdly broad, including everything from the 5 Hewlett-Packard guys in their garage to OpenAI and their billions of dollars. Especially in AI, where startups need absurd cash infusions to get off the ground, a much more useful analysis is to break them down according to the different stages of the AI-product lifecycle. The Innovation Phase starts with a vision, then the Research Phase tests out various hypotheses before the release of an AI product. The 'release' is followed by constant and repeated 'refinement' of the original model, until the AI product reaches the point of 'maturity.'

> "KY3C (which stands for "Know Your Cloud, Know Your Customer, Know Your Content") is a framework Microsoft developed to help partners and customers practice Responsible AI (and meet AI regulatory requirements). It is based on the "Know Your Customer" principle used in the financial services industry to protect against money laundering and criminal or terrorist use of financial services. Defining and assigning obligations to know one's cloud, customers, and content helps tech startups leverage other players in different layers of the AI technology stack in collectively ensuring the responsible development and deployment of AI."
>
> **Chan YOON (Director, Corporate, External & Legal Affairs at Microsoft Korea)**

2. **Tech startups often operate in different markets than "big tech," and are thus not always competitors.**

   Tech startups do not always compete directly with the 'big actors,' and only a small number may have the ambition to build a product that will directly compete with the leading firms in the ecosystem. Instead, they are competing more to produce products and services that will layer on top of existing tech products and provide better and more targeted services. This opens the possibility of sharing responsibility on ensuring that AI systems are trustworthy and collaboration between "big tech" and "little tech" to respect and promote human rights.

   Most tech startups we interviewed did not demand or expect lenient treatment from regulators. However, they asked for some patience and guidance. They admit there's a lot of catching up to do because, as compared to the more established players in the market, they are still far behind in learning how the game is played, what rules will be applied, and to what extent, etc.

3. **For AI startups, the pursuit of profits incentivizes concern for human rights.**

   An AI startup's dedication to ethics, or human rights, is often closely aligned with its profit motives. Contrary to popular belief, the startups seem to have an intuitive sense that their dedication to human rights will pay off in the long run. Because startups are so cost-sensitive, they do not engage in human rights-based service unless they have a strong conviction that the human rights-based approach will contribute to the long-term sustainability of the service and, in turn, ultimately translate into profits. This is because consumers, especially in the field of AI (perhaps because of the social anxiety associated with AI, and the aggressive ambivalence of certain tech luminaries, who travel the world speaking of the potential for greatness as well as the potential for societal apocalypse), already demand of tech startups a very high commitment to trustworthiness. There is no free pass for tech startups, and the pressure is coming from consumers as much as (or more) than from the regulators. Some startups seem to have already equipped themselves with a nuanced understanding of how their commitment to trust-

worthiness should take form, while also being open to discussions on AI guidelines.

4. **Many AI startups welcome guidance from their governments and the international community on how best to build trustworthy AI.**

   AI startups need the government to provide clarity, information, resources, and training. The business community is not automatically hostile to international normative guidance. Rather, they would regard guidance at the national or international level as helpful. As these the startups are resource constrained, the right incentives at the right time may nudge them toward legal compliance and service based in human rights. The incentives do not have to be provided in the form of monetary subsidies. Vouchers for data or compliance training would be welcomed.

5. **Tech startups have different levels of compliance needs, depending on where their pre-existing competencies have already been built up.**

   There are varying level of compliance needs and government and regulatory authorities should choose wisely when and how to intervene because they can either be very resourceful or simply be getting in the way. Startups are quite advanced in figuring out content moderation, whereas they fall behind and lack resources regarding data protection and other cybersecurity-related compliance. During the Release and Refinement Phases, startups receive constant user feedback on content moderation and community standards which forces them to reinvent their business to meet user expectations. In contrast, there is less urgency in managing privacy and cybersecurity, because user expectations only become apparent after there has been a data leakage or other mishap. These are the areas where government can make a direct and lasting impact.

6. **The "human rights" discourse fits well with how tech entrepreneurs think about managing both the risks and opportunities of AI.**

   When legal and regulatory framework around AI still seems murky at best, framing the discourse on 'what is best for our users' in the context of human rights can ultimately help guide society as to what those legal boundaries should look like. Because AI is both an empowerment tool with which we can extend the reach of our human rights agendas and a source of our deepest existential fears that the technology can degrade abilities and experiences that people consider essential being human, the 'human rights framework' can serve as an appropriate foundation upon which we can lay the building blocks for reaching humanity's full potential through AI. The term 'human rights-based obligations' may seem exceedingly burdensome when we are not even sure as to what our bare minimum 'legal obligations' are, but the universality of the doctrine could be useful especially because the legal landscape is still fluid around AI, a technology that is set to have a profound impact on the meaning of being human.

> "A human rights-based approach (HRBA) is not only a fundamental and core value but also a valuable tool in ESG practices. Utilizing the 5 steps of the UNGP human rights and business cycle can be helpful in addressing tech ethics issues. This method provides a problem-solving framework grounded in a thorough risk assessment."
>
> **Jinhwa HA (Manager of Kakao Human Rights and Tech Ethics Team)**

7. **The need for education and awareness raising. Educating and promoting human rights-oriented values across the industry will be crucial.**

   A human rights-based approach should be deeply engrained in the startup culture at least from the design/research phase. Many startups seem to display a lack of concerted efforts among the developers of AI products and the people in charge of setting up legal/ethical guardrails to infuse human rights-based considerations into product development. Before even getting at cost-benefit analysis for implementing guardrails, the idea of implementing guardrails do not even cross their minds because its importance has never been properly emphasized.

8. **Tech startups need tangible incentives and support to do what is necessary to build systems based on trustworthy AI.**

   Startups do not have the wherewithal to think deeply about these issues when they must meet launch deadlines before their funding dries up. If startups could educate policymakers on their best practices, and vice versa, and if policymakers could introduce proper incentives so that the best practices could become the industry standard rather than accelerate the startups' demise, it would be a lot easier for startups to make conscious, human rights-based decisions on their path to maturity.

9. **Leadership matters.**

   When founders to raise awareness about AI technologies, their potential benefits, and risks within the organization, other members can also voice their concerns. In turn, a culture that fosters responsible development and deployment of AI technologies will formulate organically. The idea of using 'Feed Soundness Indicators' to promote more sound content, rather than lewd, or obscene materials on Daangn's platform originated from a developer who took notice of a negative customer feedback. We witness the best practices on the ground, but the seeds of such innovation are sowed from the top reminding the constituents of the culture that is based on respect for human rights.

10. **Willingness to consider the specific context and circumstances when assessing whether an AI system adheres to a human rights-based approach is important.**

    Everyone has a different conception of what fairness is. The challenge is amplified when we unduly rely on automated decision-making through AI. Startups should aim to promote the development of AI systems that are explainable, and the governments, on the other hand, should not be so quick to judge the soundness of the entire AI system based on a selective impact assessment of those who are worse off due to the introduction of a certain innovation. Regulators should be prepared to hear out the rationale for the AI system's design before deciding that the system failed to uphold human rights altogether.

Paper 2-2:

Harnessing AI to Solve Climate Change as a "Wicked Social Problem"

AI and machine learning are often described as solutions to some of the world's most complex problems. In the medical field, for example, AI has been shown to be more accurate than (human) doctors in predicting cancer,[157] and opened the door for more precise treatment options using existing drugs on cancers of unknown origin.[158] AI similarly promises to revolutionize the field of vaccine research and design, giving rise to a whole new discipline of 'immunoinformatics'.[159] These are significant advances, often justifiably held forth as evidence of the transformative benefits of this new and still-emerging technology.

Finding a cure for cancer is a problem where a vast majority of humanity is likely to agree on the core premise of what needs to be solved, and how to go about solving it. Most people, in most communities around the world, would likely agree that a world without incurable cancer would be preferable to our current reality. In other words, there tends to be a linearity in the way most people think about curing cancer, and in automatic endorsement of experts who can credibly promise us that they hold the skillset to achieve that outcome.

Such problems lend themselves well to a science-based or engineering approach to problem solving,[160] and have therefore been described as "tame" problems. This designation says nothing about these problems' simplicity or complexity, but rather only the way these problems can be solved.[161] Many so-called "tame" problems are in fact exceedingly complex, and many have eluded human scientific and engineering ingenuity for centuries if not millennia. What makes a problem "tame" is they can be solved, it will usually happen by means of a rigorous application of the scientific method.

AI promises to enhance humanity's ability to solve such "tame" problems. AI can be thought of as a brute-force computational tool that essentially overpowers complexity by means of its sheer computational muscle. Qualcomm's Sr. Director of Engineering has described AI's ability to solve "combinatorial problems," or problems with "many choices [where researchers or planners] need to find an optimal solution."[162] Other types of combinatorial problems include supply chain management, weather forecasting, flood prediction, microchip design, and airline network planning, to name just a few. AI researchers are making significant strides towards solving such problems in ways that would have been unimaginable even just a few years ago.

This same "engineering approach" to problem solving has also been applied to many social problems, for example how to solve homelessness, the disenfranchisement of certain vulnerable populations, racism, sexism, ableism, ageism, or any number of other complex social problems we may encounter in our communities. The still-controversial philosophy of Effective Altruism (EA), popularized in the early 2000s at elite universities in the UK and the US and premised on the idea of using evidence-based approaches to maximize the positive impact of charitable efforts, is an extreme application of this problem solving-based approach towards complex social problems.

157. Zhang, Bo, Shi Huiping, Wang Hongtao, "Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: a Critical Approach," J. of Multidisciplinary Healthcare, Vol.16, 1779-1791 (2023), https://doi.org/10.2147/JMDH.S410301.

158. Moon, Intae, LoPiccolo, Jaclyn, Baca, Sylvan C. et al., "Machine Learning for Genetics-Based Classification and Treatment Response Prediction in Cancer of Unknown Primary," Nature Medicine, Vol.29, 2057–2067 (2023), https://doi.org/10.1038/s41591-023-02482-6.

159. McCaffrey, Peter, "Artificial Intelligence for Vaccine Design," In: Thomas, Sunil (ed) Vaccine Design. Methods in Molecular Biology, vol 2412. New York, NY: Humana, (2022), https://doi.org/10.1007/978-1-0716-1892-9_1; Thomas Sunil, Abraham Ann, Baldwin Jeremy, Piplani Sakshi, and Petrovsky Nilolai, "Artificial Intelligence in Vaccine and Drug Design," Methods in Molecular Biology, Vol.2410:131-146 (2022), https://doi.org/10.1007/978-1-0716-1884-4_6; and Maserat Elham, "Integration of Artificial Intelligence and CRISPR/Cas9 System for Vaccine Design," Cancer informatics, Vol.21:11769351221140102, (2022), https://doi.org/10.1177/11769351221140102.

160. Rittel, Horst and Webber Melvin, "Dilemmas in a General Theory of Planning," Policy Sciences, Vol.4(2) 155-169 (1973).

161. Id., at 158.

162. Chris Lott, "Solving Unsolvable Combination Problems with AI: How Qualcomm AI Research is Optimizing Hardware-Specific Compilers and Chip Design with AI," (Feb. 1, 2023), OnQ Blog (Qualcomm), https://www.qualcomm.com/news/onq/2023/01/solving-unsolvable-combinatorial-problems-with-ai.

The track record of the basic "engineering approach" to solving social problems has been decidedly more mixed. As Rittel & Webber observed in 1973:

The search for scientific bases for confronting problems of social policy is bound to fail, because of the nature of these problems. [. . .] Policy problems cannot be definitively described. Moreover, in a pluralistic society there is nothing like the indisputable public good; there is no objective definition of equity; policies that respond to social problems cannot be meaningfully correct or false; and it makes no sense to talk about "optimal solutions" to social problems unless severe qualifications are imposed first. Even worse, there are no "solutions" in the sense of definitive and objective answers.[163]

In their seminal essay, Rittel & Webber described what they called "wicked"[164] problems and claimed that customary science or engineering-based methods were bound to fail. Subsequent authors focused on the need for consensus building as a crucial supplement to information gathering and scientific expertise.[165] Rittel & Webber postulated ten attributes of wicked problems.[166] A subsequent analysis reduced that to three core features, claiming that they are characterized by their factual complexity, an uncertainty over how best to intervene and the precise impact(s) of those potential interventions, and finally a profound divergence of social norms on even the most foundational questions of how to conceptualize a problem.[167]

Climate change and the associated risks for the long-term survival of the human species is perhaps the most emblematic of modern-day "wicked" problems.[168] The problem of how best to fight against climate change is characterized by a staggering complexity of interconnected and poorly understood dynamics that continue to defy the best efforts of the scientific community to properly model. Since 1988, the global scientific and policy making community has attempted to reduce this complexity by pooling the collective talents of literally "[t]housands of people from all over the world" to "assess the thousands of scientific papers published each year to provide a comprehensive summary of what is known about the drivers of climate change, its impacts and future risks, and how adaptation and mitigation can reduce those risks."[169] While there is virtual unanimity among scientists that climate change is real, that it is serious, and that if left unaddressed it poses existential threats to human civilization as we know it, the specific prognostications of the Intergovernmental Panel on Climate Change remain riddled with uncertainty. This uncertainty is especially pronounced in the more granular assessments of the specific processes driving climate change or specific predictions of how global climate change might impact

163. Rittel, Horst and Melvin, supra note 161, at 155.

164. Id.

165. Alford, John and Brian W. Head, "Wicked and Less Wicked Problems: a Typology and a Contingency Framework," Policy and Society, Vol. 36:3, 397-413 (2017), https://doi.org/10.1080/14494035.2017.1361634.

166. Rittel and Webber postulated "at least ten distinguishing properties" of wicked problems: (1) there are no definitive formulation of such problems because "to understand the problem depends upon one's idea for solving it" (emphasis in the original); (2) they have no stopping rule; (3) solutions to such problems are not true-or-false, but rather good-or-bad; (4) there are no immediate and no ultimate tests of solutions to such problems, and any solution will generate "waves of consequences," the nature of which are usually difficult to anticipate and can potentially also make the problem worse; (5) every solution to such problems is a "one-shot operation", and trial-and-error attempts to find solutions are not possible; (6) such problems have no exhaustively describable set of potential solutions, nor are there parameters to the kinds of interventions one might potentially entertain; (7) wicked problem is essentially unique; (8) wicked problems can often be symptoms of another problem; (9) the choice of which explanation to select for such a problem determines the nature of the problem's solution; and (10) the planner has no right to be wrong.

167. Head, Brian, "Wicked Problems in Public Policy," Public Policy Vol.3:2, 101-118 (2008).

168. See Jim Perry, "Climate change adaptation in the world's best places: a wicked problem in need of immediate attention," Landscape and Urban Planning 133, 1-11 (2015); Frank Incropera, Climate Change: a Wicked Problem, New York: Cambridge University Press, (2016); Some have even called this a "super-wicked" problem, owing to four additional features associated with climate change. Kelly Levin and others have described also that (1) inaction on solving climate change is not cost-free, i.e., the longer one waits, the more costly it will be to solve the issue later on; (2) that those stakeholders (specifically wealthy industrialized nations) most able to solve climate change are also the ones least motivated to take action on it; (3) that there exists no global governance structure to solve what is inherently a global governance problem, and (4) that policy responses tend irrationally to discount the impacts of climate change on our collective futures. See Kelly Levin, Benjamin Cahore, Steven Bernstein and Graeme Auld, "Overcoming the tragedy of super-wicked problems: constraining our future selves to ameliorate global climate change,", 45 Policy Science 123-152 (2012); and Richard Lazarus, "Super Wicked Problems and Climate Change: Restraining the Present to Liberate the Future", 94 Cornell L. Rev. 1153-1234 (2009).

169. Intergovernmental Panel on Climate Change (IPCC), About the IPCC, (accessed Oct. 1, 2023), https://www.ipcc.ch/about/.

conditions in each locale.[170] This gives rise to a regrettable lack of direction for policy makers searching for concrete strategies to mitigate or adapt to climate change and its terrifying potential impacts.

The example of electric vehicles illustrates this conundrum. It is well understood that burning hydrocarbons in gasoline or diesel-powered personal vehicles generates significant volumes of CO2, which contributes to the greenhouse effect and accelerates climate change. An engineering approach to problem solving might suggest, therefore, that replacing such gasoline or diesel-powered vehicles with alternative "clean energy" technologies might be a handy solution to climate change. This logic would suggest investing large sums of money into ambitious research and development programs, infrastructure development, and marketing of new products to solve the problem of transportation and climate change.

At that point, however, the 'wickedness' of the climate change problem begins to complicate the analysis. A first set of questions might have to do with the complexity of the problem. Would investing in electric vehicles be more impactful than encouraging veganism, for example, given that the global meat industry arguably has a comparable or greater global carbon footprint than the entire transportation sector combined?[171] Which transition would be easier to effectuate, as a matter of costs and benefits? Debates continue to rage between well-meaning scientists and econometricians – all of whom agree that climate change is bad and needs to be prevented – over how best to compare sectors and their corresponding climate impacts. These debates are exacerbated by the vested interests of industries and sectors, for example the industrial sector with its interest in maintaining a particular business model, the labor sector with its interest in maintaining traditional means of employment, the political class interested in their own political survival, and the sociological realities of whether any of these proposed solutions are likely to be embraced by the real-world individuals who populate our society.

Debates might also rage over whether an intervention might produce the desired outcome. Will the promotion of personal electric vehicles, for example, be more impactful from a carbon footprint perspective than an equally costly alternative strategy to invest in public transportation infrastructure and integrated urban environments? Would the scramble for the natural resources necessary to produce the batteries for electric vehicles possibly plunge the world into a renewed and very costly period of geopolitical competition and conflict, and could such an outcome be avoided by simply changing our reliance on the personal vehicle as the primary mode of transportation? Such questions proliferate in the climate change debate and are often impossible to answer with definitive certainty. They also matter, however, since a choice of one strategy over another entails significant path dependencies. Getting the answer wrong is typically not something that can be easily undone, even with the benefit of perfect hindsight. Finally, the social and normative terrain surrounding these questions remains permanently contested. How would, for example, the idea of replacing privately-owned vehicles with public transportation be interpreted in the Global North, where private ownership of a vehicle has become culturally associated with prestige, power, and autonomy, as opposed to a strictly utilitarian means of moving from Point A to Point B?

This paper explores how AI can contribute to the resolution of climate change. The challenge of solving climate change may be exceptionally 'wicked,' but that has not stopped entrepreneurs, technologists, policy makers, and activists from using AI to fight climate change in its many manifestations. The Boston Consulting Group in 2023 estimated that "by scaling currently proven applications and technology, [AI] has the potential to unlock insights that could help mitigate 5% to 10% of global greenhouse gas (GHG) emissions by 2030—and significantly bolster climate-related adaptation and resilience initiatives."[172]

The entrepreneurs developing these AI use cases consider aspects of the 'wicked' problem of climate change to be inherently "solvable," even if the overarching problem of climate change remains complex. Theirs is an incremental and pragmatic approach: carving out one problem at a time and hoping that each small success will chip away at the massive mountain of a problem.

170. Pierre Gentine, "How can AI help with climate adaptation and resilience?," Learning the Earth with Artificial Intelligence and Policy (LEAP) Science and Technology Center, Columbia University, Presentation at AI for Good Workshop on The role of AI in tackling climate change & its impacts: from science to early warning, (at 1h08m), available at https://www.youtube.com/watch?v=i6yJsW3tw_4.

171. Jones, Christopher, "Cars, Cows, Coal or Consumption: Which Contributes Most to Climate Change,"(April 12, 2019), Cool Climate Network Blog, https://coolclimate.org/blog-cars-coal-cows-consumption#.

172. Amane Dannouni et. Al., "Accelerating Climate Action with AI," (Nov. 2023), BCG, https://www.gstatic.com/gumdrop/sustainability/accelerating-climate-action-ai.pdf.

This paper examines those various micro-level interventions, highlighting the specific processes that social entrepreneurs use to maximize the positive potential of their AI use cases. In other words, the paper dives into the micro-logistics of entrepreneurs using AI to "make the world a better place." For this reason, this is a paper primarily about AI as a new and emerging technology and the efforts being made by individual entrepreneurs to "make the world a better place," looking specifically at this broad field of application as a petri dish in which to examine specifically how those entrepreneurs go about doing that.

The focus of this paper diverges from a narrow focus on 'mainstream' human rights. Certainly, climate change entails important human rights considerations, specifically the right to enjoy a clean and healthy environment, and the right to life, health, and development. The overarching focus of this paper is not, however, primarily the promotion of those rights, but rather the actions that technologists, entrepreneurs, and policy makers have put in place to capture those rights-friendly upsides. What will likely become obvious, however, is the extent to which human rights considerations – especially consultation with affected communities and a concern for the impact of any given interventions on vulnerable populations, is not only necessary, but also essential for any intervention to make the world a better place to succeed.

The Urgency of Climate Change

Anyone who in 2023 is not yet convinced of the urgency of climate change is either living under a proverbial rock or willfully ignoring the preponderance of the scientific and common-sense evidence all around us. The news cycle during the summer of 2023 was dominated by wildfires, droughts, floods, heatwaves, unusually strong storm systems, melting ice sheets in Antarctica and disappearing glaciers in mountainous regions, and many more such stories too numerous to recount here. This paper will waste not additional time establishing the urgency of climate change and encourages the reader to consult the existing voluminous literature on that topic for further reference.

AI, data centers, and energy and resource consumption

At this point we should add an important caveat to this discussion. As described above, this paper looks specifically at efforts by AI entrepreneurs to capture the upsides of this new and powerful technology. In any discussion of AI and climate change, however, it must also be noted that the sheer computational power required to develop powerful AI systems require vast amounts of server power, water, and might also potentially generate substantial electronic waste in the future, once server farms start to become replaced by even more powerful hardware.[173] While these are important issues, and while many policy makers worry that on a whole, the anticipated climate-change related benefits of AI may be undone by the energy increased consumption of those powerful AI systems,[174] this paper will set aside those concerns for analytical reasons, focusing instead on efforts by AI technologists to reduce carbon output and improve the resilience of potentially affected communities.

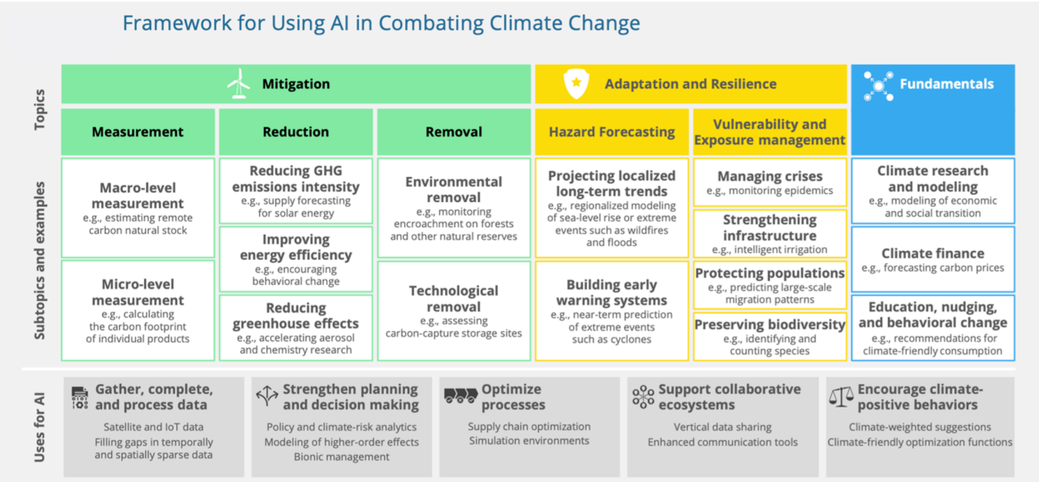
Topography of AI Applications Described as Combating Climate Change

To set the scene for this discussion, it is necessary to first survey the landscape of AI use cases that are designed, in some concrete way, to combat climate change. Some do so directly, for example by finding ways to halt the processes driving climate change, whereas others focus on the damaging impacts of climate change on ecosystems or human well-being. The first and potentially most obvious way that AI entrepreneurs have sought to combat climate change and its impacts is through the gathering of relevant data. Data serves as the catalyst allowing all other forms of action (including non-AI-enabled action) to proceed. Without it, 'evidence-based' policy making becomes impossible, and intervention campaigns begin to look more like a prolonged series of clumsy and often ideologically tainted trial-and-error efforts to solve climate change. Given the "wickedness" of climate change, such a strategy relies as much on luck as scientific skill, with massive potential dangers if the policy makers get it wrong.

173. Id. at 22-27.

174. Jim Tankersley, "The Climate Summit Embraces A.I., With Reservations," (Dec. 3, 2023), New York Times, https://www.nytimes.com/2023/12/03/climate/artificial-intelligence-climate-change.html.

Our framework builds on previous work by AI for the Planet and the Boston Consulting Group. BCG's analysis drew on scientific expertise as well as the accumulated insights of the Boston Consulting Group (BCG),[175] drawing on its various AI-related consultancies. BCG's analysis broke AI use cases into three broad fields of application: (1) efforts to assemble actionable data ("Fundamentals"), as well as AI use cases focused either on Mitigation or Adaptation efforts. Later in this paper, we argue that AI use cases might also focus on Loss & Damage efforts associated with remedying the impacts of climate change in low-resource communities. We assume that BCG's typology is non-exhaustive, and that innovators and technologists may well come up with additional use cases to combat climate change.



Source: Maher, H., Hubertus M., Damien G., et. al. (2022), How AI can be a Powerful Tool in the Fight Against Climate Change, BCG and AI for the Planet, *at* 6.

The BCG report also describes five key tasks for which AI is particularly well-suited in the fight against climate change, highlighting in particular:

1. the gathering and completing of complex data sets on emissions, climate impact, and future projections.
2. efforts to strengthen planning and decision-making processes.
3. the optimization of existing processes (for example production or recycling processes).
4. AI-enhanced support to collaborative ecosystems; and
5. the encouragement of climate-positive behaviors.

These tasks fit largely within the Predictive Analytics and Decision Support, Pattern & Anomaly Detection, and Goal-Driven Systems AI use patterns described by Cognilytica (see figure 2 above). This leaves open the possibility that future innovators will dream up additional use-cases for AI in the field of climate change building on some of the other use-patterns Cognilytica has described (for example AI applications built on autonomous systems or hyper-personalization). For the sake of consistency with the underlying model we are drawing on, we will use the 5 categories identified by BCG to drive our analysis in this paper.

175. Maher, Hamid, Hubertus Meinecke, Damien Gromier, Mateo Garcia-Novelli, and Ruth Fortmann, "How AI can be a Powerful Tool in the Fight Against Climate Change," (Jul. 2022), BCG and AI for the Planet, https://web-assets.bcg.com/ff/d7/90b70d9f405fa2b-67c8498ed39f3/ai-for-the-planet-bcg-report-july-2022.pdf.

Fundamentals

The development of better and more easily accessible scientific knowledge about climate change is a prerequisite for any efforts to effectively combat climate change. According to the BCG analysis, addressing these "fundamentals" can happen in three different ways: (1) improvements in climate research & modeling, (2) facilitated climate finance, and (3) promoting educational initiatives designed to nudge individual behavior.

Climate Research and Modeling

Climate change interventions focusing on adaptation measures often depend crucially on actionable and locally specific predictions of how climate change will impact a particular locality. Merely stating that "climate change is real," and that on average, temperature might rise by a certain number of degrees globally, has little relevance for a policy maker seeking to put in place specific adaptation measures in each municipality, region, or country. For that to happen, policy makers need more granular climate models.

AI can help generate such enhanced Climate models. AI can process vast datasets more efficiently than traditional methods and statistical models,[176] leading to more accurate and comprehensive climate models. These models can better predict weather patterns, temperature changes, and other environmental shifts. According to one expert at the European Center for Medium Range Weather Forecasting, AI-based 'nowcasting' models in 2023 are now capable of matching or even slightly outperforming the state-of-the-art statistical models used by (European) meteorologists today.[177] This pattern is true in the Global North (Europe and North America) where data sets are more historically dense and data gathering methods are systematic and multi-sourced, but not yet true for large parts of the Global South.[178]

Pierre-Philippe Mathieu from the ESA Phi-Lab explained how European data scientists were increasingly capable of cobbling together multiple sources of data into powerful comprehensive datasets, which could be used by a variety of consumers, using AI technology to pull together those various sources into one powerful tool, to serve their weather forecasting needs,[179] drawing on data from satellites, earth-based sensors, and historical records to detect even minute changes in climate variables like temperature, precipitation, and atmospheric composition. AI systems are particularly capable of managing such large datasets – datasets that would be too unwieldy for human analysts to work with – to push existing statistical models to greater granularity and precision.

Microsoft: Microsoft's so-called 'Planetary Computer' "combines a multi-petabyte catalog of global environmental data with intuitive APIs, a flexible scientific environment that allows users to answer global questions about that data, and applications that put those answers in the hands of conservation stakeholders." These resources are made available for free, and a network of developers are encouraged to create and make available applications building on the data catalog. Microsoft's 'Planetary Computer' joins other available products already available, notably Google's Earth Engine and other (paid) commercial services.

Rob Emanuele, Geospatial Architect at Microsoft who is helping to build the Planetary Computer, explains the objective of the project as "us[ing] Microsoft's technologies and capabilities to accelerate the building out of applications and solutions that have an environmental impact."[180] Microsoft's Planetary Computer project first assembles vast amounts of data and stores it on Microsoft's cloud-based storage capacities, then creates easy ways to sort and organize that data in the cloud (thereby eliminating the need to download and manipulate enormous volumes of data locally), and finally allows

176. Mariana Clare, "AI for Weather Forecasting," European Center for Medium Range Weather Forecasting (ECMWF), Presentation at AI for Good Workshop on The role of AI in tackling climate change & its impacts: from science to early warning, (at 0h15m), https://www.youtube.com/watch?v=i6yJsW3tw_4.

177. Id.

178. Id.

179. Pierre-Philippe Mathieu, "AI to Mine Datasets from Earth Observation Satellites" Presentation at AI for Good Workshop on The role of AI in tackling climate change & its impacts: from science to early warning, (at 1h15m), https://www.youtube.com/watch?v=i-6yJsW3tw_4.

180. Rob Emanuele, "The Plantary Computer," Interview on the MapScaping Podcast (with Daniel O'Donohue), https://mapscaping.com/podcast/the-planetary-computer/.

users to use open-source tools to analyze that data and integrate it into their products. According to Emanuele, "[Microsoft wants] to aim [its] capabilities at people who are making solutions that are [. . . .] used for impact."[181]

When asked about the corporate motivations for making such a resource available for free, Emanuele describes Microsoft's belief that "in order for Microsoft as a business to do well in the long term, the world needs to do well."[182] Having identified climate change as the premier threat to human well-being in our era,[183] Microsoft in 2020 made aggressive commitments to "move the needle on climate change."[184] The creation of the Planetary Computer is part of that commitment, contributing towards Microsoft's goal to promote ecosystems by using technology to promote environmental sustainability.

When asked how Microsoft intuits what kinds of data will be most valuable to users in the community, Emanuele describes a mix of a "build it and they will come" approach, mixed with input from Microsoft's corporate clients, who often request similar data and data formats for their business purposes. "Solving those [business] problems is part of our purview," says Emanuele. "The planetary computer is aimed at environmental sustainability as its primary use case, but the type of horizontal functionality that it's providing is applicable in many different use cases. Somebody who is trying to gain insights about a market based [on] changes in imagery has very similar capability needs as people who are trying to monitor land-use change to understand how carbon is changing over those areas. [T]here is a range of information that we're taking in about what are the needs for geospatial analytics, but we are also looking ahead of the puck [to ask ourselves] what is not possible right now [that we should try to make possible, and then we try to build that]."[185] (https://planetarycomputer.microsoft.com)

Climate Finance

The BCG report describes AI's role in climate finance, focusing largely on the markets for carbon credits. AI could also be used, for example, in efforts to identify new and optimized fundraising strategies in the future, especially for non-profit or humanitarian causes.

Education, Nudging and Behavioral Change

It is a well-known truism that climate change as a global phenomenon can only be addressed by the cumulative impacts of many individual changes in behavior. AI can help efforts to 'nudge' such personal behavioral choices. This can happen via chatbots such as OpenAI's ChatGPT,[186] for example, but also via a host of other conceivable personalized advice-giving tools that could be hyper-customized and individualized, the same way a personal health device might provide individualized exercise regimens to a user, for example.

181. Id.

182. Id.

183. Heather Clancy, "Microsoft is building a 'Planetary Computer' to protect biodiversity," (Apr. 16, 2020), GreenBiz Blog, https://www.greenbiz.com/article/microsoft-building-planetary-computer-protect-biodiversity.

184. Mathieu, supra note 180. (Specifically, Mathieu mentioned "announcements made around commitments that Microsoft is making around carbon, water, waste and ecosystems. Carbon: by 2030, we'll be carbon negative, and by 2050 we'll have removed all of the carbon we've ever emitted as a company since the founding; the commitment to be zero waste by 2030; to be water positive – to replenish more water than we use by 2030; and for ecosystems to by 2025 protect more land than we use and to also build a planetary computer that enables environmental sustainability solutions through the use of data in the cloud.")

185. Id.

186. Sanders, Nathan and Rose Hendricks, "AI could reshape climate communication," (Aug. 30, 2023), EOS-AGU, https://eos.org/opinions/ai-could-reshape-climate-communication.

Mitigation

Most AI use cases relating to climate change focus on mitigation efforts. According to the BCG typography, AI-enabled mitigation efforts tend to fall into three broad categories: (1) measurement, (2) carbon reduction, and (3) carbon removal. Given the increased focus on corporate practices promoting good Environmental, Social and Governance (ESG) practices, it is not surprising that many for-profit enterprises have turned to AI to help them promote such practices.

Measurement

Numerous AI-enabled tools that help educate key stakeholder groups such as business managers, investors, or consumers to 'nudge' them to make different, more climate-optimized decisions. Some of those services are developed by classical consulting firms providing specialized services to industries wishing to reduce their carbon footprints.

Eugenie.ai is a for-profit consulting company producing software as a service (SAAS) solutions for "asset heavy" manufacturers, often in the metal and mining, oil, and gas sectors. Dr. Soudip Roy Chowdhury, Founder and CEO of Eugenie, estimates that 50% of the world's greenhouse gas emissions come from just five heavy industries (oil and gas, power and utility, chemical, cement and mining), and that within those emissions approximately 10-15% are caused by process or asset inefficiency.[187] The company thus focuses its services on helping those industries identify those inefficiencies, often down to the "shopfloor" level. (https://eugenie.ai)

Persefoni is another for-profit consulting firm that works with customers to measure their carbon footprints. Persefoni's Chief Data Officer James Newsome estimates that "demand for [the company's] services could be a result of the government, investors, and shareholders putting pressure on businesses to reduce their carbon footprint" as a result of increased ESG investing practices.[188] Persefoni uses AI to analyze its clients' transactions help them reduce their carbon footprints more effectively, promising their clients the power to "manage [their] emissions data with the same rigor as [their] financial data." (https://www.persefoni.com)

Rho Impact similarly helps businesses develop their ESG strategies and potentially calculate the potential of various initiatives to help cut their carbon footprint. The company has even proposed developing an open-source AI-powered tool to help investors make better investment decisions using AI-generated data about publicly listed companies' efforts to tackle climate problems.[189] (https://rhoimpact.com)

Reduction

Most AI-enabled climate change mitigation efforts focused on efforts to reduce carbon emissions. These efforts typically focus on capturing and eliminating inefficiencies, often by means of AI systems designed either to reduce the intensity of Greenhouse Gas (GHG) emissions, improve energy efficiency, or drive scientific advances that may seek to alter the basic chemistry driving climate change in some way.

DeepMind Technologies Ltd: DeepMind was founded in 2010 and acquired in 2014 by Google. In 2023 DeepMind was merged with Google Brain to "bring together two of the world's leading AI labs"[190] and presumably also to compete more effectively with emerging AI rivals . DeepMind rose to fame by designing AI programs trained using a deep reinforcement learning approach that could successively outcompete human players at various video games. Learning only from the pixels on a screen, various iterations of DeepMind AIs eventually beat the world's top-ranked players playing some of the most technically complex human games such as AlphaGo and other games.

187. Dr. Soudip Roy Chowdhury, "Interview with Fotis Georgiadis," (Jun, 12, 2022), Authority Magazine, https://medium.com/authority-magazine/eugenie-ai-dr-soudip-roy-chowdhurys-big-idea-that-might-change-the-world-d558ef4b69b1.

188. Aaron Mok "Some climate-tech startups want you to believe AI tools can save the planet–but it's not that simple," (May. 24, 2023), Business Insider, https://www.businessinsider.com/can-ai-be-a-solution-to-the-climate-crisis-2023-5.

189. Fischer, I., Beswick, C., & Newell, S., "Rho AI–Leveraging artificial intelligence to address climate change: Financing, implementation and ethics," Journal of Information Technology Teaching Cases, 11(2), 110-116, at 112 (2021).

190. Google DeepMind, "Our mission" (accessed Nov. 29, 2023), https://deepmind.google/about/.

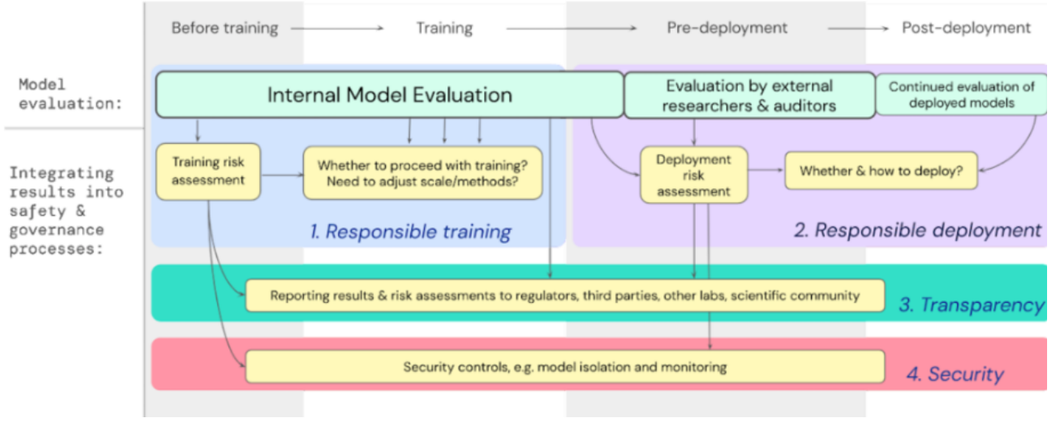
These demonstrations of the technologies' potential translated to a host of practical use cases, for example in medicine where the technology was able to accurately predict protein folding (a previously unsolvable problem), various text-to-speech systems which are now integrated into common consumer electronics products and coding applications. DeepMind also has a long history of collaboration with the UK health services to use AI in the detection of early onset blindness, or cancer research.

In 2017, after DeepMind was acquired by Google but before it was merged with Google's own in-house AI research lab, Deep mind created a new research unit called DeepMind Ethics & Society. This unit brought together "experts from the humanities, social sciences and beyond, along with voices from civil society and technical insights from [the DeepMind team] to conduct and fund interdisciplinary research."[191]). Early on, this process developed a set of 7 principles Google DeepMind will use to "develop technology responsibly."[192]

1. Be socially beneficial.
2. Avoid creating or reinforcing unfair bias.
3. Be built and tested for safety.
4. Be accountable to people.
5. Incorporate privacy design principles.
6. Uphold high standards of scientific excellence.
7. Be made available for uses that accord with these principles.

Owing to the institutional heft of Google, one of the most iconic tech-giants in the field of AI, these principles have become universally well-known and well-regarded.

The research lab also developed a process for identifying and guarding against novel threats that is particularly relevant as technologists begin to approach general purpose AI models, as opposed to the currently-available narrow AI models.[193] This process seeks to "expand the evaluation portfolio to include the possibility of extreme risks from general-purpose AI models that have strong skills in manipulation, deception, cyber-offense, or other dangerous capabilities" (emphasis in the original). This model mirrors the risk-based regulatory approach to AI and asks responsible AI developers to "look ahead and anticipate future possible developments and novel risks."[194] This analysis is not limited only to risks inherent to the model itself, but also the risks of nefarious or negligent use of an AI by less-than-ethical bad actors. Using this model, Google DeepMind focuses on the responsible training, responsible deployment, transparency, and appropriate security of its higher-risk AI products. Notably, this approach is broken down into tangible processes, similar to the HRBA@Tech model, that translate the approach into concrete activities that Google DeepMind scientists and engineers can do to ensure the responsible development and deployment of their technologies.



191. Sean Legassick and Verity Harding, "Why we launched DeepMind Ethics & Society" (Blog post of Oct. 3, 2017), https://deepmind.google/discover/blog/why-we-launched-deepmind-ethics-society/.

192. Google AI, "Our Principles," https://ai.google/responsibility/principles/.

193. Toby Shevlane, "An early warning system for novel AI risks," (Blog post of May 25, 2023), https://deepmind.google/discover/blog/an-early-warning-system-for-novel-ai-risks/.

194. Id.

As one of the world's premier AI research labs, Google DeepMind's ambitions are expansive, going far beyond the issue of climate change. Nonetheless, DeepMind is often cited as an organization that has developed innovative solutions to climate change. Using its deep reinforcement learning method, the company was able to reduce the energy consumption of its data centers by 40%.[195] Given that over 40% of the overall energy consumption of a typical data center goes to keep the computer servers cool,[196] this represents a significant reduction in these data centers' climate impact. Just as it did to "play games", DeepMind's RL system learned to explore safe configurations for the server farms that had not previously been explored, leading to non-intuitive discoveries that significantly boosted energy efficiency, such as spreading loads across more equipment. DeepMind used different mechanisms to ensure the system would still behave as intended, such as verifying optimal actions against an internal list of safety constraints defined by data center operators, who retained the option at any time to regain control over the server's operations.

Using this same approach, Google DeepMind reputedly entered initial talks in 2017 with the UK's National Grid to optimize the grid's energy usage without adding any new infrastructure.[197] In 2019, it applied a similar approach to better predict the time-based electricity output of wind farms in the United States, thus boosting their "value" to the power grid by 20%.[198] (https://deepmind.google/about)

Focusing on the potential of companies to make the world a better place by promoting responsible AI often encounters the obvious concern about profitability. Perhaps it is to be expected that Google – a multi-billion-dollar global tech giant – would invest resources into the odd AI-use case with no obvious commercial benefit. As one commentator put it when describing Google DeepMinds' application of its RL technology to wind farms in the United States:

[DeepMind] has done phenomenal work from a research perspective, but has yet to find substantial revenue streams. It loses a lot of money ($368 million in 2017), which has reportedly contributed to tensions between DeepMind and [Google]. If the company's software can be put to use in real-life scenarios outside the research lab, DeepMind could become a revenue-generating segment of the business that justifies its high costs.[199]

Other stakeholders, notably national governments, are less beholden to the raw logic of profit when they contemplate investments in pro-social technology and can therefore be counted on to incur some of the startup costs of spurring innovation, subsequently promoting those innovations for use by the private sector.

Korean Smart Grid: The Korean government similarly invested to produce "smart grids," similar to the collaboration described above between Google's DeepMind and the UK Power Grid. In Korea, however, the initiative was spearheaded by the government.

In 2009 the island of Jeju hosted the smart grids demonstration project, which lasted until 2013. The project validated 153 relevant technologies and distilled nine potential business models that made smart grids commercially viable. As a result of this demonstration project, most private companies in Korea are incorporating smart grid technologies such as electric vehicle charging systems and demand response (DR) aggregator models (that operate by consolidating consumer demand based on predictive models and temporarily store that energy in Energy Storage Systems). The Korean government captured this learning in the Korea Smart Grid institute (K-SGI), which is mandated to spread the use of smart grid technology throughout the country. The Korea SGI participated in the Jeju smart grids demonstration project and has also supported other similar use cases. One example is the Advanced Metering Infrastructure (AMI) dissemination and big data platform project, which distributed over 2.45 million smart meters to households across the country between 2020 to 2022, thus greatly enhancing the capacity for smart grids to work. In another example, the K-SGI supported a demonstration project in Sumi Village in Yangpyeong (Gyeonggi Province) to prove

195. Richard Evans & Jim Gao, "DeepMind AI Reduces Google Data Centre Cooling Bill by 40%," (Jul. 20, 2016), Google DeepMind blog, https://deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/.

196. Zhang, X, T. Lindberg, N. Xiong, V. Vyatkin and A. Mousavi, "Cooling Energy Consumption Investigation of Data Center IT Room with Vertical Placed Server,", 105 Energy Procedia, 2047-2052 (2017).

197. Madhumita Murgia, Nathalie Thomas, "DeepMind and national Grid in AI talks to balance energy supply," (Mar. 12, 2017), Financial Times, https://www.ft.com/content/27c8aea0-06a9-11e7-97d1-5e720a26771b.

198. Nich Statt, "Google and DeepMind are using AI to predict the energy output of wind farms," (Feb. 26, 2019), The Verge, https://www.theverge.com/2019/2/26/18241632/google-deepmind-wind-farm-ai-machine-learning-green-energy-efficiency.

199. Id.

the feasibility of an effort to cobble together various small renewable energy sources to replace larger (and fossil-fuel dependent) power plants.

The Korean government's efforts to centralize (and fund) these demonstration projects has served as an effective accelerator of efforts to reduce carbon emissions in Korea, both by increasing the efficiency of the power grid and by making renewable energy more readily available. In line with its commitment to progressively realize social and economic rights (which, unlike in the case of the corporate sector is consistent with a government's 'business model'), the Korean government has also focused on extending the benefits of smart grid technology to people in lower-income and underprivileged areas, who are often expected to be the most severely impacted by rising heating and cooling costs. In an ongoing effort to promote energy self-sufficiency, Korea Southern Power has recently installed solar panels in households headed by persons living with disability in the city of Busan.[200] If such initiatives continue, the socially vulnerable populations will also benefit from the expansion of smart-grid technology in Korea in ways that might not be possible if the effort were left solely to the free market to resolve.

Other AI climate-related use cases focus on efforts to improve energy efficiency, either by reducing the use of energy or finding better and more efficient ways to reduce, reuse and recycle existing resources.

Mortar io: Existing buildings contribute approximately 40% of global carbon emissions, and 80% of the buildings that are likely to exist in 2050 are already standing today.[201] According to its founders, these datapoints justify a focus on the decarbonization of existing structures as one of the primary climate change mitigation challenges. This small UK-based startup, which in 2022 had no more than 3 employees, is at the pre-seed fundraising phase. During its first few months of existence, the company "secured partnerships with some of the largest property management companies in the UK [and was hired to] model and simulate over 150,000 buildings for a London borough."[202] These early commercial contracts helped the company field-test the technology to promote net-zero emissions for existing buildings.

The company offers "physics-based modelling, enriched by AI to create retrofit twins (a concept based on the notion of a digital twin) for commercial buildings. [These digital twins allow building portfolio managers] to model, plan, and execute retrofit projects for hundreds of buildings simultaneously, [enabling them] to make informed decisions in minutes rather than months."[203] (https://www.mapmortar.io/)

Arup (Neuron): Arup is a legacy engineering firm, founded in the 1940s, that enabled iconic engineering projects such as the Sydney Opera House, the Centre Pompidou in Paris, the HSBC building in Hong Kong, the Øresund Bridge linking Denmark and Sweden, and the National Aquatics "Water Cube" designed for the 2008 Beijing Olympics. Developed initially in the context of an innovative construction engineering project in Hong Kong, Arup later supported the development and commercialization of Neuron as an independent company.[204] In 2022 the company partnered with Venturous, a "city-tech" group, to "leverage the power of data and technology to decarbonize building assets and facilitate the transformation towards digital property management."[205]

Neuron provides its clients with an intuitive and fully customizable visualization tool that enhances buildings' energy savings, improves efficiency, and optimizes operational workflows. Neuron uses 5G and Internet of Things (IoT) sensors to gather real-time data from building equipment and systems, and then uses AI to optimize HVAC operations, support air quality improvements, and provide insights on effective building carbon reduction strategies.[206] These innovations have the potential to save between 10-30% of the energy consumed by a typical commercial building.[207] (https://www.arup.com/services/digital/arup-neuron), (https://www.neuroncloud.ai/)

Recycling, which is widely touted as a contribution towards efforts to mitigate the climate crisis, is also being impacted by AI technologies. That is especially true for metals. Recycling aluminum,[208] for example, reduces CO2 emissions by 92% compared to using new aluminum. The figure for recycled steel and copper is 58% and 65% reduction in CO2 emissions respectively.[209] That said, separating these metals from regular (non-recyclable) waste often poses significant challenges, and frequently results in otherwise recyclable materials being dumped in a landfill.

Zen Robotics (a Finland-based company) and AMP Robotics (a US-based company): Both companies use AI to more efficiently identify and sort recyclable waste. Zen Robotics claims to be the first company to apply AI-based robots to waste sorting in 2009.[210] It was acquired in 2022 by Terex, a major industrial machine manufacturer.[211] AMP robotics was founded in 2014, and by 2019 had raised $16 million in Series A funding in 2019, another $55 million in Series B funding in 2020, and a further $91 million in Series C funding in 2022.[212] Both of these companies are thus at the very tail end of what we might describe the Product Lifecyle associated with startups.

Tim Dewey-Mattie, Recycling and Public Outreach Manager at Napa Recycling in California, describes how this technology helped them be more efficient identifying recyclables that are "actually worth money" to them, not to mention the value of those items not going to the landfill. "We're always looking for information, different stats and different numbers about how [our recycling operation] is going. [This system] gives us a lot of data that we didn't necessarily have before."[213] Dewey-Mattie also describes how these technologies have allowed them to reply to their human workers "into jobs that are really great for human brains."[214] (https://www.terex.com/zenrobotics/), (https://www.amprobotics.com/)

Refiberd: Refiberd is a California-based different recycling company focusing on recycling fabrics. 60 percent of textile materials are today made of plastic, largely fueled by the "fast fashion" industry, which churns out approximately 200 billion new items of clothing annually (enough for every person on the planet to consume an average of 25 new items of clothing each year). Over 80% of those clothes are eventually discarded in landfills or incinerated,[215] including about half a million tons of microplastics, most of which end up being washed back into the ecosystem.[216] The industry churns out more CO2 than the aviation and shipping industries combined and consumes 93 billion cubic meters of water per year.[217]

The company was founded by a women-led team of engineers during the pandemic. The recycling system makes use of AI, near-infrared radiation spectroscopy and robotics-based recycling engineering.[218] The technology is designed to recycle materials, especially fabric, that have proven to be difficult to distinguish in conventional recycling processes.[219] The company offers integrated solutions for existing conveyor belts, but also

200. Sang Bok, "Southern Power, installing solar modules o the roofs of vulnerable," (Mar. 7, 2023), E2 News (Korean language), https://www.e2news.com/news/articleView.html?idxno=251472.

201. Mortar io, "About Mortar io," (Dec. 1, 2023), https://www.mapmortar.io/.

202. Mortar io, "Mortar IO secures preseed funding!" (Blog post of Oct. 16, 2023), https://www.mapmortar.io/post/mortar-io-secures-preseed-funding.

203. Id.

204. Arup, "Ventures", (accessed Dec. 1, 2023), https://www.arup.com/our-firm/ventures.

205. Jerman Cheung, "Arup and Venturous Group launched Neuron Digital Group in a quest to make buildings smarter," (Arup webpage, accessed Dec. 1, 2023), https://www.arup.com/news-and-events/arup-and-venturous-group-jointly-established-neuron-digital-group.

206. Global Partnership on AI, Climate Change and AI: Recommendations for Government Action, 87 (2021), https://www.gpai.ai/projects/climate-change-and-ai.pdf.

207. Id.

208. EuRIC AISBL - Recycling: Bridging Circular Economy & Climate Policy, "Metal Recycling Factsheet," https://circulareconomy.europa.eu/platform/sites/default/files/euric_metal_recycling_factsheet.pdf.

209. Id.

210. Zenrobotics, "Our History," (accessed Dec. 1, 2023), https://www.terex.com/zenrobotics/about-us/our-history.

211. Id.

212. AMP Robotics, "Origins," (accessed Dec. 1, 2023), https://www.amprobotics.com/origins.

213. AMP Robotics, "Napa Recycling & Waste Services," (Blog post of Jul. 19, 2022), (see embedded YouTube video), https://www.amprobotics.com/casestudies/napa-recycling-waste-services.

214. Id.

215. Refiberd, "Intelligent sorting for textile-to-textile recycling," (accessed Dec. 2, 2023), https://refiberd.com/.

216. Owen Mulhern, "The 10 Essential Fast Fashion Statistics," (Jul. 24, 2022), Earth.org (blog post), https://earth.org/fast-fashion-statistics/.

217. Id.

218. Refiberd, "About Us," (accessed Dec. 1, 2023), https://refiberd.com/about/.

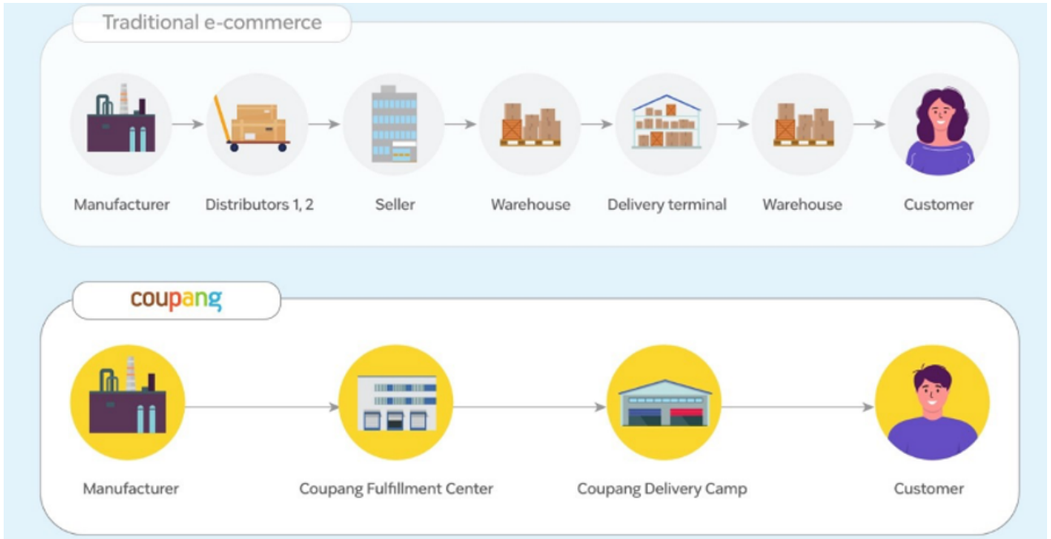
219. Sarika Bajaj, "Refiberd: Textile Waste Sorting for Recycling," (Apr. 27, 2023), Presentation at the 27th Annual Recycling Update - 2023, (accessed Dec. 1, 2023), https://youtu.be/8OAVXrU5QcA?si=wrbnfwfgYrtR8tLT.

provides remote sorting capabilities for mechanical and chemical recyclers.[220]

The two founders of the startup envisioned a future where their technologies might end up recycling up to 93% of worldwide fabrics. Their AI-powered approach to circular fashion received early recognition and financial support from several accelerators and news agencies. The company won numerous awards, such as the SXSW Pitch, Fashion for Good, and an award from the H&M Foundation, helping them attract notoriety and raise $3.4 million in funding to support their mission.[221] (https://refiberd.com/about/)

Other companies embrace environmental protection as a means of improving their economic efficiency and corporate reputation, and then discover–almost as an unexpected byproduct– new business models associated with AI-enabled sustainable business practices.

Coupang: Coupang is a major Korean online marketplace, often referred to as the "Amazon of Korea." Coupang went public in 2021, thus officially 'graduating' from our definition of a startup company. Coupang strives to reduce its carbon footprint by using AI technologies to minimize the number of steps along the supply chain connecting the manufacturer with the consumer. Conventional e-commerce platforms deliver products to customers using a chain of intermediary distributors, inevitably emitting CO2 and generating waste along each node in that sequence. Coupang's AI optimization strategy has created an "end-to-end" system whereby it manages every step from the product manufacturer to the delivery to the customer. AI fuels this optimization process making it faster, more efficient, and eco-friendlier.



(Source: Coupang Newsroom)

Coupang's commitment to focus on building an eco-friendlier system is part of its corporate commitment to ESG goals. Kang Han-Seung, CEO of Coupang, believes that Coupang's turn to technological innovation broke the tradeoff between customer convenience and sustainability.[222]

Coupang's innovations are driven first and foremost by a focus on economic efficiency. Coupang's AI system evaluates countless variables, spanning inventory choices, product locations, delivery routes, and external factors such as the predicted weather and the geographical terrain to optimize delivery strategies. These innovations have also resulted in a 50% reduction of food wastage in Coupang's inventory and a staggering reduction of 30,000 tons of Styrofoam packaging annually. Courtesy of these efficiencies, Coupang was also able to develop entirely new business models, for example the rapid delivery of so-called "ugly" vegetables or ripe seasonal fruits and vegetables from farmers directly to consumers, thus unlocking a revenue stream for Coupang (and farmers) that previously would not have been possible through an online marketplace. (https://www.aboutcoupang.com/.)

AI can also help drive scientific advances that then open innovative business opportunities for entrepreneurs intent on making environmental sustainability part of their business mandate.

Solidia Technologies and Uncountable: Solidia Technologies is a Texas-based cement startup, founded in 2007, to "develop and commercialize technology with CO2 mineralization technology."[223] CO2 mineralization injects carbon into cement mixes, making the cement stronger and permanently mineralizing the carbon (which was previously captured from the cement factory's smokestack) directly into future batches of cement. According to the company's reporting, this can reduce concrete's carbon intensity by up to 70%.[224]

In 2020, Solidia Technologies partnered with a Uncountable, a group of data scientist engineers, to take advantage of its data analysis expertise to "expedite R&D and applications of Solidia's next-generation, sustainable concrete manufacturing process."[225] Solidia CEO Tom Schuler mentioned that "Uncountable helps us accelerate product development by dramatically speeding data analysis, helping us predict iterations and move much more quickly."[226] According to Schuler, "[n]ever before has the industry had [. . .] a means of rapid-fire testing that can expedite production upgrades and efficiencies, new recipes, and improved performance in concrete." This AI-enhanced analytical capacity allows the company to expedite the evaluation of new chemical formulations of concrete to meet the specifications of clients while also dramatically reducing the product's overall carbon footprint."[227] As Uncountable's Founder Noel Hollingsworth put it, his company's "AI algorithms [. . .] mitigate[] the need for tedious, manual tweaking of individual ingredients. Complicated development processes that would involve tens or hundreds of experiments are now conducted by Solidia in half the time utilizing advanced machine learning models."[228] (https://www.solidiatech.com), (https://www.uncountable.com)

Removal

Should efforts to reduce climate emissions fail, other scientists are working on solutions to physically remove existing carbon from the atmosphere and store them safely (and permanently) in other formats. Without a doubt, the most effective current-day technology to remove carbon is nature itself, primarily in the form of forests and the oceans, that remove carbon from the environment. This can be called environmental removal, since these processes are central to the world's natural carbon cycle. Scientists have also been working to develop artificial processes to remove carbon (technological removal). Tech entrepreneurs have been finding ways to support both environmental and technological removal strategies with AI.

Environmental Removal

Several projects have combined AI with satellite and other remote sensing technologies to help predict, monitor, and respond to threats to forests and healthy biospheres.

Rainforest Connection: The Rainforest Connection has developed a "Guardian Platform" that operates devices in the tree canopies of rainforests equipped with high-capacity microphones. These devices connect to satellites and upload the recordings to a cloud-based server, where it is analyzed for threat identifiers, such as the sound of chainsaws, etc. The system uses AI to "deliver rapid insight into what's happening in the vast forest ecosystems" being monitored, "identify poten-

220. Id.

221. Alexandra Harrell, "$3.4M Funding is 'Just a Tool' for this Textile Recycling Startup," (Sep. 6, 2023), Sourcing Journal, https://sourcingjournal.com/sustainability/sustainability-news/refiberd-funding-textile-recycling-startup-artifical-intelligence-sarika-bajaj-452927/.

222. Coupang, ""Building the future of e-commerce": YTN Science reports on Coupang" (citing to a Korean-language report on YouTube, https://youtu.be/tkcqeNkS9FQ?si=9Y8cqHC4SdcuFGFr), (May 18, 2022), https://www.aboutcoupang.com/English/news/news-details/2022/Building-the-future-of-e-commerce-YTN-Science-reports-on-Coupang/default.aspx.

223. Solidia, "About Us," (accessed Dec. 1, 2023), https://www.solidiatech.com/about-us/.

224. Solidia, "Cement and concrete technology engineered for greener businesses," (accessed Dec. 1, 2023), https://www.solidiatech.com/technology/.

225. Solidia, "New enterprise software platform brings data revolution to cement and concrete," (Blog post of Jan. 8, 2020), (accessed Dec. 1, 2023), https://www.solidiatech.com/solidia-technologies-and-uncountable-accelerate-development-of-next-generation-concrete-manufacturing-with-data-science-and-ai/.

226. Id.

227. Id.

228. Id.

tially harmful behavior, and help on-the-ground rangers pinpoint and stop damaging activities as they occur."[229] (https://rfcx.org/guardian)

Harvard University (AI for Conservation): The Harvard John A. Paulson School of Engineering and Applied Sciences runs a research program called AI for conservation. In it, researchers focus on developing AI use cases to "wildlife protection and the protection of natural resources." One AI-enhanced project (the Protection Assistant for Wildlife Security, or PAWS) helps park rangers in developing nations plan more effective anti-poaching raids based on previously collected data about poaching patterns. PAWS uses machine learning to predict criminal behavior, and generates risk-models for conservation officials to plan their patrol routes.[230] The same project also seeks to establish a strategically located checkpoints around conservation areas to catch smugglers who seek to export illicit wildlife products to the global market.[231] A second project uses Unmanned Aerial Vehicles (UAVs, or "drones") equipped with an AI-enabled "SPOT" tool designed to identify the thermal infrared signatures of suspected poachers operating in a nature reserve.[232] (https://teamcore.seas.harvard.edu/ai-conservation)

Google Earth Engine: In a precursor to Microsoft's "planetary computer" (see above), Google's Earth Engine has become an invaluable resource to conservationists. In 2013, this system was able to produce the first-ever map of the world's forests and how they had changed since the turn of the millennium – a feat that would have taken a single computer 15 years to compute.[233] Other use cases include an interactive map for conservationists to view habitat ranges of individual animal species, a tool to "measure and visualize changes to the world's forests," a tiger habitat monitoring system, a malaria risk mapping map, and an "open-source tool to visualize and analyze plots of land" to assess deforestation and land-use changes."[234] (https://earthengine.google.com/)

Technological Removal

A growing field of climate change science has to do with efforts to capture carbon before it is released into the atmosphere, or to harvest it back out from the atmosphere. According to one recent study on the issue, "by optimizing factors including temperature, pressure, flow rates, and chemical reactions, AI algorithms can be used to improve the efficiency and effectiveness of CO2 capture processes."[235]

Xyonix: Xyonix is an AI consulting firm offering several specialized services. One service promises to 'build custom AI and machine learning models to improve [Carbon Capture and Storage Systems, or CSS systems] to more effectively mitigate climate change.[236] The company describes how AI systems can be trained to discover optimal absorbents that maximize the separate carbon molecules from the atmosphere.[237] The company notes how Total Energies, IBM, and Haliburton are all currently using this AI-enabled technology to develop more efficient commercial strategies for mechanical carbon storage. (https://www.xyonix.com)

229. Rainforest Connection, "Guardian Platform," (accessed Dec. 2, 2023), https://rfcx.org/guardian.

230. Harvard John A. Paulson School of Engineering, "PAWS: Protection Assistant for Wildlife Security," (accessed Dec. 2, 2023), https://teamcore.seas.harvard.edu/paws-protection-assistant-wildlife-security.

231. Harvard John A. Paulson School of Engineering, "Illegal Smuggling and Global Wildlife Trade Prevention," (accessed Dec. 2, 2023), https://teamcore.seas.harvard.edu/illegal-smuggling-and-global-wildlife-trade.

232. Harvard John A. Paulson School of Engineering, "Machine Learning for Wildlife Conservation with UAVs," (accessed Dec. 2, 2023), https://teamcore.seas.harvard.edu/machine-learning-wildlife-conservation-uavs.

233. Jéssica Maes, "How Google Earth Engine revolutionized the way we monitor deforestation," (Jun. 8, 2023), The Verge (accessed Dec. 2, 2023), https://www.theverge.com/23746844/google-earth-engine-amazon-deforestation-monitoring.

234. Google Earth Engine "Case Studies," (accessed Dec. 2, 2023), https://earthengine.google.com/case_studies/.

235. Priya, A.K., Balaji Devarajan, Avinash Alagumalai, and Hua Song, "Artificial intelligence enabled carbon capture: A review," (2023), Science of the Total Environment 886, 163913.

236. Xyonix, "AI in Carbon Capture and Sequestration," (accessed Dec. 2, 2023), https://www.xyonix.com/industries/environment/carbon-sequestration.

237. Mackenzie Komeshak, "Using AI to Optimize Mechanical Carbon Capture & Storage systems," (Blog post of Feb. 15, 2022), (accessed Dec. 2, 2023), https://www.xyonix.com/blog/using-ai-to-optimize-mechanical-carbon-capture-amp-storage-systems.

Adaptation & Resilience

Scientists and entrepreneurs are also devoting considerable efforts to help societies transition to a "new normal" that accepts – at least for a transitional period – that societies will have to learn to cope with a changed climate. This can be thought of as the "Plan B" to successfully mitigating the impacts of climate change.

The BCG report breaks Adaptation and Resilience focused AI use cases into two broad categories: (1) Hazard Forecasting and (2) Vulnerability and Exposure Management. With some minor exceptions for corporations seeking to protect their assets, most of these applications tend to fall more of a humanitarian or development framework and tend therefore to be less driven by corporate profit motives than the mitigation efforts described above (many of which – when successful – qualify for ESG investment and carbon credits). Jonas Weiss, from IBM Research, admitted as much when he acknowledged that "it's a big challenge to create a business model with these applications."[238]

Thus, while the potential role of AI remains strong, also in Adaptation and Resilience efforts, the business case for developing these use cases remains weaker.

Hazard Forecasting

To prepare communities to prepare for a 'new normal,' scientists and policy planners need good models to predict the long-term trends. The problem, however, is that current models are relatively poor at predicting with any specificity how the climate will change at the local or regional level.[239]

AI systems currently are much more able to accurately predict imminent threats. These early warning systems contribute significantly to local resilience. Pierre Gentine (Columbia University) highlighted efforts by Google, FloodBase (a Brooklyn startup) to use satellite data to map flood risk and inundations in near-real time. This can help with emergency response and evacuation efforts during natural disasters, such as hurricanes or flash flooding.[240] Other researchers are working to develop monitoring systems that can track the spread of wildfires in near-real time.[241]

Overstory AI: One company that has built a business case attempting to predict hazards is Overstory AI. This consultancy uses satellite data and AI to predict the likelihood of certain natural objects like trees or shrubbery to impact critical infrastructure. The company markets these services to the operators of critical infrastructure, for example electricity grids or railroad companies. (https://www.overstory.com)

Vulnerability and Exposure Management

Moving even further into the domain of humanitarian relief, some relief agencies are also beginning to deploy AI in efforts to minimize vulnerability to climate-fueled disasters. The BCG typology lists four sub-categories of such vulnerability management: managing crises, strengthening infrastructure, protecting populations, and finally preserving biodiversity.

The UN World Food Program, for example, has moved to a near real-time strategy for monitoring food insecurity. In 35 countries, this model depends on continuous data collection (primarily through live-phone calls to affected areas), and in an additional 50+ countries it depends on a "nowcasting" model that draws on a series of input data such as prevailing market prices, the number of conflict-related fatalities in conflict zones,

238. Jonas Weiss, IBM Research "Large Scale AI Models to Create Earth Observable Actionable Insights," (Sep. 25, 2023), Presentation at AI for Good Workshop on the role of AI in tackling climate change & its impacts: from science to early warning, https://www.youtube.com/watch?v=i6yJsW3tw_4.

239. Pierre Gentine, "How can AI help with climate adaptation and resilience?," (Sep. 25, 2023), Learning the Earth with Artificial Intelligence and Policy (LEAP) Science and Technology Center, Columbia University, Presentation at AI for Good Workshop on the role of AI in tackling climate change & its impacts: from science to early warning, (at 1h08m), https://www.youtube.com/watch?v=i6yJsW3tw_4.

240. Id.

241. Ban, Y., Zhang, P., Nascetti, A. Et al., "Near Real-Time Wildfire Progression Monitoring with Sentinel-1 SAR Time Series and Deep Learning," (2020), Sci Rep 10, 1322, https://doi.org/10.1038/s41598-019-56967-x.

rainfall data, etc.[242] This model can be used to predict food insecurity emergencies with a 30-day horizon with only a 4% error rate.

Consistent with what one might expect of an organization whose explicit mandate is to "make the world a better place", the WFP emphasizes the need for such models to take into consideration the local specificities of a particular intervention. As Giulia Martini explained it, "Expertise from the field is crucial" and "extensive consultation with stakeholders is imperative."[243] Such consultations are important not only to train and verify the AI system itself, but also to develop local confidence in the integrity of the data and predictions. Furthermore, such intensive cultivation can be crucial to capture local specificities that may not be apparent in more 'universalized' approaches, for example the importance of Ramadan in a food security context, where global models may not account for such locally relevant cultural practices.[244]

Similar efforts are underway to begin predicting the risks of internal displacement, tracking not only the relationship between weather events on displacement, but also other factors such as the socio-economic status of a household as a predictive factor in that family's displacement.[245]

Other initiatives are designed specifically to help vulnerable communities build their resilience in the face of a changing climate.

Conservation International (CI) is an American nonprofit environmental organization. In partnership with Arizona State University in the United States, Konservasi Indonesia (a national foundation promoting sustainable development in Indonesia) and Thinking Machines Data Science (a tech consultancy focused on Southeast Asia), CI initiated the "Climate Smart Shrimp" (CSS) Program in Indonesia and the Philippines to help address severe mangrove deforestation. Traditional patterns of shrimp aquaculture have historically been one of the primary factors of global-level-deforestation and ecosystem destruction.

CI initiated the CSS program to support the Global Mangrove Alliance's 2030 goal of increasing mangrove cover globally by 20%. The CSS program attempts to advance this goal by working to restructure one of the greatest threats to mangroves today, the aquaculture sector, and to do so in a way that diminish environmental harm, facilitates ecosystem restoration, attracts investment capital, and scales up restoration initiatives.[246] It does so by intensifying the shrimp farming process while also combining it with a restoration of mangrove along the coastline. This combined effort increases the yield of shrimp farmed, while also vastly improving the biodiversity of the coastal areas and reducing harmful runoff into the ocean ecosystem. The restored mangrove forests also help to protect vulnerable coastal communities in the face intensifying climate-related natural disasters.

In subsequent iterations of the same program (in Ecuador), CI combined this program with a supply chain certification process that allowed buyers to purchase shrimp from only those farms participating in the program, using those funds to sustain the overall initiative.[247]

To succeed, CI needed to identify suitable areas for this model to work. Relevant factors include "proximity to roads and populated areas, proximity to historical and present mangroves, pond size, and slope and elevation." To make these assessments at scale, CI worked with partners to develop an AI-powered Multi-Criteria Decision Analysis (MCDA) approach to identify over 40,000 suitable hectares where this model might work across Southeast Asia. Those were then the areas where traditional development practitioners went as they interacted with communities, conducted follow-up assessments, and worked to establish individual programs to promote shrimp farm conversion efforts.[248]

Once a suitable location has been identified, the CSS Program consists of two separate initiatives. The first provides technical assistance to communities and farmers. CI Aquaculture experts investigate local factors such as the water temperature, oxygen levels and salinity to understand how to optimize conditions for shrimp growth and health, while aerators were introduced to ensure optimal oxygen levels in the artificial ponds hosting the shrimp aquaculture. In addition, the CSS Program extends loans to shrimp farmers to fund their intensification and restoration initiatives, allowing the farmers to upgrade their farm equipment and infrastructure.

While CI initiated this project, it actively consults with farmers, supply chain companies, government, investors, communities, local NGOs, and other stakeholders throughout the process. The objective of these collaborations is to find locally sustainable ways to restore blue carbon ecosystems and to decrease the environmental impacts of shrimp farms, while also enhancing the productivity and resilience of shrimp farms. (https://www.conservation.org/)

Loss & Damage

The BCG report was published prior to the conclusion of the 27th session of the Conference of the Parties to the UN Framework Convention on Climate Change in Sharm el-Sheikh, Egypt (COP27), which took place in late 2022. At that conference, the global community finally agreed to create a Loss and Damage Fund to compensate communities for irreparable harms caused by climate change. The Fund was established one year later at the 2023 COP28 in Dubai. Loss and Damage has been a topic of significant concern for decades, especially in the Global South, where communities often struggle under the financial and social toll of mounting climate-change related disasters. The issue has long been taboo among more wealthy industrialized nations, fearful that "Loss and Damage" would become tantamount to an admission of guilt by the industrialized North for climate change. Not comfortable with such straightforward notions of accountability for the harms associated with climate change, many industrialized nations argued that Loss & Damage should be subordinated into discussions of adaptation and resilience efforts.

Now that Loss & Damage has been recognized as a third pillar of efforts to combat climate change, it makes sense to expand the BCG framework to consider how AI can contribute also to those important efforts. Loss & Damage is still an evolving concept, but it is already possible to speculate that AI will also play an important role in efforts to compensate and support communities in the Global South for the harms caused – whether recklessly or unknowingly – by the economic activities of the generally more industrialized and wealthy societies of the Global North.

Loss & Damage can be understood to cover two basic scenarios. In the first, a community is severely impacted by the effects of climate change such that a recovery is desirable but impossible, given the limited resources available for such a reconstruction effort. In such situations, any meager development progress that may have been achieved under 'climate normal' conditions are quickly reversed by successive waves of climate-change fueled disasters. Entire communities can be relegated into perpetually worsening cycles of poverty, indebtedness, and human despair. Loss & Damage would demand that a concerted effort be mounted by wealthier industrialized nations (which, coincidentally, are also primarily responsible for the historical GHG emissions that gave rise to climate change in the first place) to support those communities as they rebuild sustainably and with greater resilience than before to the ongoing and intensifying climate hazards they face. This is not development assistance by another name, but rather a rights-based intervention designed to allow those vulnerable communities to exit a cycle of crisis survival and refocus on a more dignified and sustainable development trajectory.

Investing in such a strategy would require innovative financing vehicles that move beyond an ad-hoc (disaster-by-disaster) approach to humanitarian aid and lackluster rebuilding efforts. It would require the design and

242. Giulia Martini, UN World Food Programme "Early Warning for Food Systems," (Sep. 25, 2023), Presentation at AI for Good Workshop on the role of AI in tackling climate change & its impacts: from science to early warning, https://www.youtube.com/watch?v=i-6yJsW3tw_4.,

243. Id.

244. Id.

245. José María Tárraga Habas, "Explainable AI and Causal ML for Disaster-Induced Displacement," (Sep. 25, 2023), Presentation at AI for Good Workshop on the role of AI in tackling climate change & its impacts: from science to early warning, https://www.youtube.com/watch?v=i6yJsW3tw_4.

246. Anica Araneta, "Scaling Climate Smart Shrimp in Southeast Asia using GIS and Computer Vision," (Jun. 16, 2022), Climate Change AI Blog Post, https://www.climatechange.ai/blog/2022-06-16-grants-mangrove.

247. CI, "Conservation International and xpertSea Launch "Climate Smart Shrimp" Regenerative Farming Pilot in Ecuador," (May 23, 2023), https://www.conservation.org/press-releases/2023/05/25/conservation-international-and-xpertsea-launch-climate-smart-shrimp-regenerative-farming-pilot-in-ecuador.

248. Araneta, supra note 247.

mobilization of innovative construction strategies, preferably ones that don't seek merely to rebuild old infrastructure, but rather improve it and build in a more resilient and sustainable way. Here too, AI can play a significant design and research role. Loss & Damage programming would also require new and more effective ways of extending the benefits of insurance to vulnerable communities, the same way that insurance spreads risk in more socio-economically secure parts of the world. Finally, AI could be used to rapidly identify communities and areas that qualify for such assistance, and develop scalable accountability mechanisms for funds distributed, in ways that do not needlessly deplete funds with overhead costs and notoriously sluggish assessment, verification, and inspection regimes relying exclusively on human development experts.

Microsoft's AI for Good Research Lab actively undertakes projects relating to climate change. These projects focus on the goals of loss and damage and adaptation. The research lab oversees a range of different projects.

Under its AI for Good Research Lab, Microsoft established an African AI Innovation Council "to harness the power of data and AI to boost climate resilience and adaptation efforts in Africa."[249] The Innovation Council will convene members from leading African regional governance organizations such as the African Development Bank, African Risk Capacity, and the African Climate Foundation. The Council will work to identify opportunities to improve climate resilience through data and AI and facilitate ways to generate additional climate data and drive continued research.

Juan Lavista Ferres, Chief Scientist and Lab Director at Microsoft's AI for Good Research Lab noted "the indispensable need for collaboration [. . .] throughout the Council's discussions,"[250] quoting an African proverb to illustrate the point: "If you want to go fast, go alone. If you want to go far, go together." Lavista Ferres continued that "[t]his is why the AI for Good Lab remains committed to our partnership model, working with subject matter experts to guide the responsible development of AI solutions so that they are maximally useful and accessible to all."[251]

Microsoft also runs a collaboration with Planet Labs to apply AI technology and satellite data to support African climate adaptation projects.[252] This is especially crucial due to the climate data divide whereby there is insufficient reliable climate data and a lack of data scientists to work with the available data to turn them into insights for decision-making. By combining Planet Labs' high-quality satellite imagery of Africa with Microsoft's AI technology, Africa-based data scientists will have access to satellite imagery from across the African continent to develop adaptation strategies and early warning systems. (https://blogs.microsoft.com/on-the-issues/2022/11/07/climate-data-divide-global-south/)

The second Loss & Damage scenario is that of a community that simply becomes no longer viable, either for an individual household or for an entire community. The paradigmatic example would be a low-lying island nation in the Pacific Ocean threatened by rising sea levels, or a farming household in Sub-Saharan Africa or Central America driven by prolonged droughts and famine to migrate to Europe. What would happen in such scenarios, where staying home is simply no longer a survivable option? One recent study found that "over the coming 50 years, 1 to 3 billion people are project to be left outside the climate conditions that have served humanity well over the past 6,000 years" and that "absent climate mitigation or migration, a substantial part of humanity will be exposed to mean annual temperatures warmer than nearly anywhere today."[253]

A range of new challenges arise in such a scenario. The first has to do with combatting the sheer human indignity of such situations. Being forced to leave one's home is always traumatic, regardless of whether it is driven by climate change, conflict, poverty, or other reasons. All kinds of displacement are traumatic, but the fact that millions of people are displaced annually does not render this phenomenon as mere background noise: regrettable but too big to solve. Quite the contrary, forced migration remains one of the most urgent sources of human suffering globally.

Some of this indignity has to do with the legal framework that governs contemporary migration flows. Existing international and national legal frameworks do not classify individuals and communities displaced by slow-onset climate change events as refugees. Climate migrants lack the legal protections afforded to families fleeing war or even those affected by a rapid-onset natural disaster such as an earthquake or tsunami. Border guards typically categorize climate-migrants as "illegal" economic migrants, refusing them entry and leaving them with no choice but to go back home, where survival is impossible, or place themselves at the mercy of the exploitative and often violent criminal trafficking networks who can smuggle them across international borders.

This simply cannot be an acceptable baseline "solution" for countries, communities, and households reeling from the irreparable impacts of disastrous Loss & Damage. The global community can and must do something. But it is also realistic to expect that policy makers will want to find some way to distinguish between climate-displaced migrants and "economic migrants." In so doing, governments will likely turn to AI, as they already have at most international borders, to help with the smooth identification and transfer process. These border initiatives will likely need to be coupled with well-funded political and humanitarian initiatives to support climate migrants, to streamline migratory flows, and to provide legal and dignified alternatives to the human trafficking networks.

Beyond merely moving people to new locations, Loss & Damage would also require efforts to preserve the cultures, identities, and memories that define displaced communities. A person always has the right to exercise his or her culture, language, and traditional practices, even after being displaced from their former community. Technologies such as language translation and AI-enabled VR/XR technologies can help preserve and maintain cultural practices even when communities disperse or move from their original homelands. This can happen in cultural or religious contexts, but also for educational authorities wishing to continue educating children about the language and cultural traditions of the "homeland."

Finally, any such scenario would raise the question of what would happen to the political institutions that govern a community. Displaced communities still retain the right to preserve their identity and their political heritage. In some situations, this has implications for national sovereignty, for example when a country simply disappears off the face of the earth. What is left of a country if there is no longer any physical territory to stand on? Can there be a governance model that continues to have genuine meaning for the former "citizens" of that polity, and can this be facilitated by AI technologies? There are numerous tech utopians who would gladly design innovative and exciting AI-facilitated governance models to serve precisely such purposes. Indeed, the low-lying small island nation of Tuvalu has famously created a digital twin of the entire country in anticipation of the potential extinction of the physical landmass of the country due to climate change.[254]

Our expanded framework of AI efforts to combat climate change, building on the BCG model but adding a focus also on Loss & Damage, would look as follows:

Amended Framework for using AI to Combat Climate Change (including Loss & Damage)

Mitigation			Adaptation and Resilience		Loss and Damage		Fundamentals
Measurement	Reduction	Removal	Hazard Forecasting	Vulnerability and Exposure Management	Repair & Rebuild	Human Dignity, Identity and Culture	
Macro-level measurement e.g., estimating remote carbon natural stock	**Reducing GHG emissions intensity** e.g., supply forecasting for solar energy	**Environmental removal** e.g., monitoring encroachment on forests and other natural resources	**Projecting localized long-term trends** e.g., regionalized modeling of sea-level or extreme events such as wildfires and floods	**Managing Crises** e.g., monitoring epidemics	**Insurance & Appraisal** e.g., risk management strategies	**Governance** e.g., virtual communities and e-government	**Climate research and modeling** e.g., modeling of economic and social transition
	Improving energy efficiency e.g., encouraging behavioral change			**Strengthening infrastructure** e.g., intelligent irrigation	**Construction** e.g., build back better, temporary shelter & economic transition management	**Migration** e.g., planning and regularization of climate-induced migration flows	**Climate finance** e.g., forecasting carbon prices
Micro-level measurement e.g., calculating the carbon footprint of individual products	**Reducing greenhouse effects** e.g., accelerating aerosol and chemistry research	**Technological removal** e.g., assessing carbon-capture storage sites	**Building early warning systems** e.g., near-term prediction of extreme events such as cyclones	**Protecting populations** e.g., predicting large-scale migration patterns	**Accountability** e.g., anti-corruption monitoring & traceability	**Culture and Memory** e.g., digital strategies to preserve and maintain cultural heritage	**Education, nudging, and behavioral change** e.g., recommendations for climate-friendly consumption
				Preserving biodiversity e.g., identifying and counting species			

249. Sylvester Addo, "COP27 – Microsoft announces new Africa AI Innovation Council," (Blog post of Nov. 10, 2022), https://microsoft-caregh.com/2022/11/10/cop27-microsoft-climate-africa-ai-innovation-council/.

250. Juan M. Lavista Ferres, LinkedIn Post, https://www.linkedin.com/posts/jlavista_this-week-in-nairobi-microsoft-joined-with-activity-7091444689068838912-1DDk?utm_source=share&utm_medium=member_desktop.

251. Id.

252. Ayooluwa Adetola, "Microsoft and Planet Partner to Provide AI and Satellite Data for African Climate Adaptation Projects," (Nov. 17, 2022), Space in Africa, https://africanews.space/microsoft-and-planet-partner-to-provide-ai-and-satellite-data-for-african-climate-adaptation-projects/.

253. Chi Xu, Timothy Kohler, Timothy Lenton, Jens-Christian Svenning and Martin Scheffer, "Future of the Human Climate Niche," 117(21) PNAS, 11350-11355 (2020), https://doi.org/10.1073/pnas.1910114117.

254. Kalolaine Fainu, "Facing extinction, Tuvalu considers the digital clone of a country," (Jun, 27, 2023), The Guardian, https://www.theguardian.com/world/2023/jun/27/tuvalu-climate-crisis-rising-sea-levels-pacific-island-nation-country-digital-clone.

Analysis: Integrated Lessons Learned from the Case Studies

Five lessons stand out from the above analysis:

Making the world a better place deserves the world's brightest minds.

The Google Deepmind principles highlighted above make this point when they state that a commitment to the responsible use of technology requires a company to "uphold high standards of scientific excellence." At one level, this is perhaps obvious, especially in a for-profit model where technology is being developed and deployed to paying customers. But as one moves along the spectrum of AI use cases designed to make the world a better place, there can be an understandable temptation on the part of some technologists to embrace technologies that are merely "good enough." This tendency can sometimes (regrettably) be seen in any pro-bono activities, where professionals volunteer their expertise and services for free or at below-market prices. Compromising on the quality of those products or services can be harmful for several reasons, not the least of which being that AI products, even when specifically designed to meet socially beneficial purposes, can still pose significant risks to users and communities.

This serves as an important reminder to technologists working to design products that make the world a better place. Noble as those intentions may be, this does not free them from still needing to also focus on the potential downsides of novel technologies, such as the harms that a poorly designed AI system could still have on communities.

The principle also recalls an age-old problem facing any non-for-profit or public service venture, namely how to incentivize talented individuals with highly marketable skillsets to choose less lucrative, but potentially more socially beneficial, career options. Corporations, governments, universities, and non-profit organizations need to continue finding creative ways to share talent, not just with the highest bidders, but also with the most socially deserving causes.

"Big tech" has a significant role to play.

The second realization is that "big tech" has a significant role to play in "making the world a better place." Many of the innovations highlighted above come from major tech corporations – Google and Microsoft, but also many others that we did not mention but could easily have also highlighted, including Meta, IBM, NVIDIA, etc. Cynics may describe these efforts as corporate "greenwashing" or window dressing – flashy efforts designed to cover up for an otherwise unremarkable track record on climate change or human rights issues. And in some cases, such a cynical perspective is certainly warranted.

That said, not all efforts by major corporations to 'make the world a better place' are necessarily motivated by some nefarious motivation. Many of those highlighted above, certainly, are not. What our analysis makes clear is that without these private corporations dedicating talent to conduct meaningful safety research, consult with communities and users, fund special projects like Google Earth or Microsoft's Planetary Computer, global efforts to fight climate change would be in a much worse place than they currently are. As difficult as this may be for some climate activists to accept, especially those who see the often-times negative impact that corporations have had on the environment or on human rights, it must be recognized that not all corporations are cut from the same cloth. Furthermore, those corporations that do, in good faith, mobilize resources to make the world a better place should be encouraged and recognized for their efforts. This happens, for example, quite prominently during the annual COP Climate Change Conferences. At one level, this acknowledgment may come from the positive public relations impact of pro-social activities. But civil society and government also has a role to play to reinforce that positive feedback loop, rather than only always focus on the negative reinforcement mechanisms associated with naming and shaming (for civil society groups) and regulation and criminal prosecution (for government authorities).

NGOs must get smarter to help them scale their efforts with AI and other new and emerging technologies.

The Conservation International case study shows the importance of traditional non-profit organizations and civil society groups also turning to AI, or at least partner organizations who know how to make value of AI, to 'scale' and improve their own services. The act of "helping" and empowering communities is also in need of being turbo-charged by AI technologies. Non-profits have a responsibility to inform themselves of the tremendous potential for AI, and to build their own capacities, however challenging that may be, to take maximum advantage of those technologies.

There is a thriving consulting industry making use of AI technologies to help them advise for-profit businesses on how to implement better ESG practices, but most of those efforts focus primarily on the "E" of ESG.

Quite a few case studies highlighted above describe for-profit consultancies offering to use innovative and AI-enhanced to advise their clients on improving their ESG practices. The market for such services is obviously thriving and perhaps growing. One realization, however, is that many of those AI use cases focus on helping corporations become more environmentally sustainable. Few, so far, focus on social or governance sustainability. While beyond the scope of this chapter which focused only on efforts to combat climate change, AI entrepreneurs should also develop ways to advise corporations on how to reduce their social and governance impact on surrounding communities and use AI to enhance those efforts.

Efforts remain primarily "top down" in their approach. There are only few examples where any significant consultation took place with affected communities, and those examples tended to come from government or development / humanitarian agencies familiar with the need for rigorous consultation processes.

Finally, and perhaps most strikingly, we note that many of the efforts described above still fail meaningfully to engage with impacted communities. The crucial exceptions to this pattern tend to be civil society organizations and governments, both of which tend to be incentivized and well-practiced at the art and science of "consulting a vulnerable community." This consultation process requires some effort, but it is also central to any genuine effort to 'make the world a better place.'

Corporations certainly know how to structure such processes. Corporate advertising departments virtually invented the practice of convening focus groups to "test" a product or message with key demographics. Business consultants are well-versed in terms like 'stakeholder mapping' and 'consensus building.' Especially when it comes to AI, they need to more visibly and more aggressively liaise with potentially impacted stakeholders to ensure that their voices are reflected in any efforts – especially those oriented towards 'making the world a better place.' Moreover, they need to create non-intimidating and viable grievance mechanisms such that stakeholders can come directly and at an early stage to the company or tech entrepreneur to highlight negative impacts a technology may be having.

Paper 2-3:

The Global Governance Landscape of AI and its Potential to Better Promote a Human Rights-Based Approach to AI

This paper transitions from the specifics of how AI works and how AI engineers and entrepreneurs are working to mitigate the downside risks of AI while also capturing its tremendous potential upsides and refocuses on the policy making arena. The chapter details the efforts of policy makers at the corporate, civil society, national and international level to influence how AI is governed, and concludes with some recommendations specific to the United Nations on some additional reforms that could make the UN—and in particular its Human Rights Institutional Machinery—more able to also engage constructively with this ongoing discussion.

The Rapidly Evolving AI Policy Landscape and the Need for a Human Rights–Based Approach

Corporate self-regulation and the proliferation of AI principles

AI has only recently exploded into popular conscience with the mass-commercialization of large language models, such as OpenAI's Chat GPT. Corporations, on the other hand, have been experimenting behind closed doors with these technologies for years, and have also been promulgating policies to guide their own internal processes and ensure the safety of their AI products. The Chinese tech giant Tencent, for example, published its AI principles in collaboration with the Chinese Academy of Sciences as early as 2017, and Google and Microsoft produced their AI principles in 2018.[255] These primarily principle-based frameworks make up a panoply of overlapping ethical norms.

There have also been some industry-wide efforts to coordinate such self-regulation efforts. The Partnership on AI, for example, brings together researchers from more than fifty of the biggest American tech companies to collaborate on AI ethics and governance. It serves as a research and information sharing platform and includes practical tools such as the AI Incident Database that documents failures of AI systems around the world.[256]

Though AI start-ups do not always publish such internal safety guidelines or principles, they often integrate ethical commitments prominently into their mission statements. Corporate executives increasingly make public statements or commitments about the importance of 'responsible' or 'safe' AI systems. VCs, who have traditionally been quite hesitant to embrace ESC criteria into their investment strategies, are also gradually starting to think about industry-wide efforts to ensure that new AI-use cases are still trustworthy. In November of 2023, for example, over forty of the largest venture capitalist firms – collectively managing hundreds of billions of dollars that fund AI start-ups – signed a non-binding charter of Responsible AI Commitments[257] aimed at providing specific guidance on the development of responsible AI. Such commitments will become increasingly impactful if funding for AI start-up continues its exponential growth.[258]

One meta-study analyzing various proposed governance frameworks on AI found a gradual convergence towards a set of 8 core principles articulated in these various frameworks:[259]

1. **Privacy.** "AI systems should respect individuals' privacy, both in the use of data for the development of technological systems and by providing impacted people with agency over their data and decisions made with it."
2. **Accountability.** There should be "mechanisms to ensure that accountability for the impacts of AI systems is appropriately distributed, and that adequate remedies are provided."
3. **Safety and Security.** "AI systems [should] be safe, performing as intended, and also secure, resistant to being compromised by unauthorized parties."
4. **Transparency and Explainability.** "AI systems [should] be designed and implemented to allow for oversight, including through translation of their operations into intelligible outputs and the provision of information about where, when, and how they are being used."
5. **Fairness and Non-discrimination.** "AI systems [should] be designed and used to maximize fairness and promote inclusivity."
6. **Human Control of Technology.** "Important decisions [should] remain subject to human review."
7. **Professional Responsibility.** Individuals play a vital role "in the development and deployment of AI systems [and should consider it as their professional duty to ensure] that the appropriate stakeholders are consulted, and long-term effects are planned for."
8. **Promotion of Human Values.** "The ends to which AI is devoted, and the means by which it is implemented, should correspond with our core values and generally promote humanity's well- being."

These broad areas of ethical concern encompass stringent and detailed principled guidance on what should be included to realize these overarching objectives. They also all remain inherently subjective. In the words of Philip Alston, UN Special Rapporteur on Extreme Poverty, "as long as you are focused on ethics, it's mine against yours. I will define fairness, what is transparency, what is accountability. There are no universal standards."

With regards to the need to protect privacy, for example, numerous ethical codes point to the importance of seeking permission from users by means of notice-and-consent regimes,[260] while others point to a supposedly higher standard of seeking informed consent.[261] Similarly, while many codes stipulate that users should have meaningful control over the use of their data, this idea can be interpreted in various ways. Microsoft, for example, commits itself to 'appropriate controls' on how data is used, whereas IBM's principles stress that 'users should always maintain control over what data is being used and in what context.'[262]

Such privacy safeguards can vary from mere commitments to provide information on how data is used, to concrete procedures designed to allow users can change, restrict or delete personal data from being used.

In the same vein, while almost all AI principles contain some form of commitment to the principles of fairness and non-discrimination, this also covers a range of different measures in pursuit of those principles. Some codes contain commitments that AI-systems' training data will be representative of the target population. Others contain commitments that outputs will not replicate or amplify existing social discriminations. Some commit the firm to ensure that users will be treated impartially and equitably, while others promote equal opportunity and a proactive effort to correct for societal inequalities, ensuring inclusiveness and representativeness by empowering historically marginalized populations.[263] This diversity of specific interpretations of what is required by "fairness and non-discrimination" evidence a mix of both "do-no-harm" provisions as well as some focusing on actively "making the world a better place."

Despite the differences in how various ethics codes interpret largely similar-sounding principles, the true test of a company's commitment to an abstract principle (privacy, fairness, etc...) is not so much in its willingness to publicly embrace those principles but rather its commitment to put in place concrete mechanisms to breathe life into those principle. A study of the 24 companies with the highest standards of AI principles found that 'more than half of the committed firms had not introduced any concrete steps towards responsible AI' by introducing internal or external mechanisms to monitor compliance.[264] One example of a company that did put in place such mechanisms is Microsoft's AI Ethics in Engineering and Research program. This initiative brings together 'experts in key areas of responsible AI, engineering leadership, and representatives nominated

255. Field, Jessica, et. Al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI," (Jan. 15, 2020), Berkman Klein Center for Internet and Society Resarch Publication Series No. 2020-1, https://cyber.harvard.edu/publication/2020/principled-ai.

256. Partnership on AI, "AI Incidents Database," https://partnershiponai.org/workstream/ai-incidents-database/.

257. Responsible AI Commitments for startups and their investors, https://www.rilabs.org/responsibleai-commitments.

258. VC funding for generative AI startups went from $3.9 billion in 2022 to $17.8 billion in 2023. See Raquel Jorge Ricart and Pau Álvarez-Aragonés, "The geopolitics of Generative AI: international implications and the role of the European Union," (Nov. 27, 2023), https://www.realinstitutoelcano.org/en/work-document/the-geopolitics-of-generative-ai-international-implications-and-the-role-of-the-european-union/?utm_source=pocket_saves.

259. Field, Jessica, et. al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI,"(Jan. 15, 2020), Berkman Klein Center for Internet and Society Research Publication Series No. 2020-1, https://cyber.harvard.edu/publication/2020/principled-ai.

260. Sundar Pichai, "AI at Google: Our Principles," (Jun. 7, 2018), https://www.blog.google/technology/ai/ai-principles; A Latam, "Declaración de Principios Éticos Para La IA de Latinoamérica," (2019), http://ia-latam.com/etica-ia-latam/.

261. IBM, "IBM Everyday Ethics for AI," (2019), https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

262. Microsoft, "AI Principles," 68 (2018), https://www.microsoft.com/en-us/ai/our-approach-to-ai; IBM, "IBM Everyday Ethics for AI,"44 (2019), https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

263. Microsoft, "AI Principles," 69 (2018), https://www.microsoft.com/en-us/ai/our-approach-to-ai; Artificial Intelligence Industry Alliance, "Artificial Intelligence Industry Code of Conduct," (2019), (See Principle 3), https://www.secrss.com/articles/11099; University of Montreal, "Montreal Declaration for a Responsible Development of Artificial Intelligence,"14 (2018) (See Principle 7.4), https://www.montrealdeclaration-responsibleai.com/the-declaration.

264. Laat, P., Companies Committed to Responsible AI: From Principles towards Implementation and Regulation?, Philos. Technol. 34, 1135–1193 (2021), https://doi.org/10.1007/s13347-021-00474-3.

by leaders from major divisions.'[265] These representatives are divided into several Working Groups, focusing on subthemes such as 'AI Bias and Fairness'. The Committee is collectively mandated to work with product teams to ensure that Microsoft's products align with the company's AI principles.

Regulatory Responses and State-led Initiatives on AI

As the risks associated with AI have become more apparent, starting in 2016, policy makers have also begun to explore a range of potential regulatory responses. This has not been a straightforward process, however. From the outset, efforts to regulate AI were in tension with counterarguments that regulation would stifle innovation and competition, and potentially harm national tech industries vis-à-vis their counterparts operating in less-regulated jurisdictions. Until recently, this was illustrated by the European Union's cross-sectoral, risk-based regulatory approach, which was often contrasted with the United States' relatively liberal (i.e., hands-off), market-driven and sectorally-targeted approach towards regulation.

Domestic AI regulation can adopt essentially two forms. It can have a broad scope, addressing the implications of AI in a comprehensive, cross-sectoral manner, or it could be targeted at specific uses or risks arising from AI applications. The latter approach seems to be taking hold in more countries. Globally, since 2016, around 31 countries have adopted a total of 123 AI-related bills.[266] The US leads in the enactment of these laws, having adopted 22 texts, between 2016 and 2022, followed by Portugal and Spain, which have adopted thirteen and ten laws respectively.[267] This, however, does not encompass broad AI legislation, but bills that contain provisions on the use of AI for targeted areas such as training programs, public administration decision-making, or the impacts of AI on education.

Targeted regulation can allow for greater flexibility to address specific concerns over certain human rights risks or processes. The US Data Privacy Protection Act, for example, was introduced in May 2022 and contains provisions requiring large data holders to conduct impact assessments of their algorithms if these pose risks of harm to individuals or groups.[268] Targeted regulations can also be more easily adaptable in response to advances in technology. China, for example, has adopted regulations on recommendation algorithms and on synthetically generated content.[269] Such targeted regulation also, however, requires governments to essentially play a game of 'whack-a-mole': constantly enacting new and specific regulations to address new problems as they arise. This is complicated, especially considering the lightning pace of AI evolution and the near-instant impact of any AI-systems gone wrong. Therefore, while a reliance on targeted regulations is certainly appropriate (and even necessary in certain instances), a stronger and more overarching approach may be necessary as a complementary measure to guide the spirit of those more targeted interventions.

The US approach, which has so far followed the country's preference for market development, innovation, and competition, is gradually shifting towards a more regulatory-favorable position. In an October 2023 Executive Order, President Biden outlined the US' intention to lead in the context of rapid advancement of AI capabilities 'for the sake of our security, economy, and society.'[270] This recent push to define a unified US regulatory approach is an effort by the Biden administration to put in place an alternative to the regulatory models emerging in Europe, China and other markets, and to reassure US-based tech companies that they can continue to drive AI innovation and experimentation in the United States. The recent Executive Order sets out eight principles to guide the use of AI: (1) safety and security; (2) responsible innovation, competition, and collaboration; (3) support for workers; (4) equity and civil rights; (5) consumer protections; (6) privacy and civil liberties; (7) risk management in the Government's own use of AI; and (8) leadership. The Order is broad in that it discourages general bans on the use of generative AI, leaning instead towards limited access to specific AI services based on risk assessments.

In contrast to targeted approaches, broad-based regulation, like the proposed EU AI Act, can bring a comprehensive, risk-based approach to AI. The EU AI Act broadly defines AI as software that can "generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with."[271] The Act, which is expected to become binding law in early 2024, forms part of the EU's regulatory approach to AI. It focuses on the rights of users and citizens, and would form a broad spectrum protection of the fundamental rights and democratic structures, while working to promote a fair distribution of AI's benefits.[272] A prominent feature of the Act is that it ensures blanket protections for different levels of risk. For AI systems that pose high risks to fundamental rights, health and safety, the Act requires high-quality data, documentation and traceability, transparency, and oversight. Going a step further, the proposed act would also outlaw the use of real-time facial recognition in public places (due to its risk of compromising fundamental rights and freedoms), as well as other AI systems that manipulate human behavior or exploit individuals' vulnerabilities.[273] For AI systems considered to be non-high-risk, the Act requires companies to implement codes of conduct, thereby ensuring a minimum level of protection for users interacting with these systems. In preparing the draft AI Act, the European Commission assessed alternative regulatory strategies, including a legislative instrument with a voluntary labeling scheme, an ad hoc sectoral approach, a horizontal legislative instrument following a risk-based approach, and a horizontal legislative instrument establishing mandatory requirements for all AI systems. The Commission's ultimate choice to put forward a horizontal legislative instrument relying on a two-tiered risk-based approach, while mandating codes of conduct for lower-risk systems, represents a hybrid option between these various regulatory approaches.

A draft of the EU AI Act has prompted some noteworthy push-back, primarily from politicians in France, Germany and Italy.[274] Concerned about the impact of this EU AI Act on the competitiveness of their own tech sectors working with AI, these efforts proposed for there to be a greater reliance on industry self-regulation. Such an approach, which has been dubbed a more "innovation friendly" model, could potentially dilute the regulatory 'bite' of the proposed EU AI Act. The French Minister for Economy, Finance, and Industrial Sovereignty has stated that "before regulating AI, the EU must innovate" if it wants to remain in the 21st Century race for AI.[275]

Broad regulations cover a wide range of AI applications, and yet they also have drawbacks. Certain emerging or unanticipated issues can easily fall through the cracks in such regulatory regimes. Emerging human rights concerns may, in fact, be better addressed using specific regulation. The best approach may therefore be to enact broad regulation while also allowing for specific and more nimble regulation where necessary.

China's regulatory approach strikes a balance etween State control and encouraging innovation and competition. China began to develop regulation ahead of many other countries, establishing a timeframe in 2017 to guide national regulation efforts, in an attempt to lead the regulatory race over its main competitors.[276] Most recently, China enacted the world's first law addressing generative AI, with comprehensive requirements that training data used by AI companies must be 'true and accurate,' and prohibiting racial and gender discrimination. These regulations have sparked dis-

265. Microsoft, "The Microsoft Responsible AI Standard, v2: General Requirements," (Jun. 2022), https://www.microsoft.com/en-us/ai/principles-and-approach?activetab=pivot1%3aprimaryr5.

266. Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault, "The AI Index 2023 Annual Report," (Apr. 2023), AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf.

267. Id., at 269.

268. H.R.8152 – American Data Privacy and Protection Act (117th Congress, 2021-2022) https://www.congress.gov/bill/117th-congress/house-bill/8152/text.

269. Matt Sheehan, "China's AI Regulations and How they Get Made," (Jul. 10, 2023), Carnegie Endowment for International Peace, https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117.

270. Executive Office of the President, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," (Nov. 1, 2023), 88 Federal Register 75191, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

271. European Commission, "Proposal for a Regulation of the European parliament and of the Council Laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," (Apr. 21, 2021), Brussels, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206, Article 3: Definitions

272. Id.

273. Id. at Art. 5(1)(d).

274. Will Henshall, "EU's AI Regulation Could be Softened After Pushback from Biggest Members," (Nov. 22, 2023), Time Magazine, https://time.com/6338602/eu-ai-regulation-foundation-models/.

275. Agence France Presse, "Bruno Le Maire: L'Union Européenne Doit "Innover Avant de Réguler l'IA", (Nov. 5, 2023), CBNews, https://www.cbnews.fr/digital/image-bruno-maire-union-europeenne-doit-innover-avant-reguler-ia-80054.

276. Sarah Zheng, Jane Zhang, "China Wants to Regulate its Artificial Intelligence Sector without Crushing it," (Aug. 14, 2023), Time Magazine, https://time.com/6304831/china-ai-regulations/.

cussions within China as to whether and how legislation may thwart innovation and competition.[277]

In parallel to (or in addition to) the more slow-moving regulatory efforts to ensure the safety of AI systems, many states have also pursued various soft law solutions, ranging from multi-stakeholder processes to sandboxing, issuing codes of conduct, and the issuance of general guidance. These various efforts focus not only on the development of new technologies, but also the corporate strategies used to bring them to the market. This was also the approach of the EU until 2021, before its shift to a more regulatory approach. In April of 2019, the EU's High Level Working Group on AI published its Ethics Guidelines for Trustworthy AI, and in June of that same year it published its Policy and Investment Recommendations for Trustworthy AI, in which it for the first time proposed the risk-based approach to AI regulation that the European Union would later embrace in the form of the EU AI Act. Many other countries, including Colombia, the Republic of Korea and the United Kingdom, have also adopted national AI strategies. Still other countries, including India and the Philippines, have established or proposed establishing specialized government agencies tasked with examining the need for regulation and standardization of AI technologies.

Other State-led initiatives have taken place at the intergovernmental level, in fora like the G7 and the G20, the World Economic Forum, the OSCE, and UNESCO, to name just a few. The G7, for example, began to consider AI regulations at an ICT ministerial meeting in 2016. Since then, it has developed a set of commitments to promote 'a human-centric AI'[278] and the Global Partnership on Artificial Intelligence (GPAI), which fosters a 'vision of AI that is human-centered, fair, equitable, inclusive and respectful of human rights and democracy.'[279] After years of negotiation, the GPAI was finally published in 2019.[280]

The G20 adopted a similar approach when in 2019 it adopted a set of principles focused on human-centered AI.[281] These were reaffirmed by heads of State in September of 2023, formalizing those governments' commitment to 'pursue a pro-innovation regulatory/governance approach' maximizing the benefits and considering the risks associated with AI.[282]

There have also been several legislative efforts to control how States themselves use AI for state functions. Governments are increasingly turning to AI for a range of uses, including improving healthcare and other social services, conducting traffic flow analysis, tracking undocumented migrants and improving border control measures, and a host of other government functions.

Policy makers in some countries have shown an appetite for regulating governmental use of AI, balancing a desire to promote government innovation with an ongoing concern for human rights. In November 2023, for example, the US Office of Management and Budget, following the aforementioned executive order issued by President Biden, released guidance on the use and development of AI by the government to 'provide a model for the responsible use of the technology.'[283] The draft guidance recommends Chief AI Officers for federal departments to advise and track government AI activities, expand reporting on the ways government agencies use AI, and mandate the implementation of specific safeguards for AI uses that impact rights and safety.[284] 31 UN member States also recently signed a declaration on the responsible military use of AI and autonomy, recognizing the need to develop military AI in a responsible and ethical way that enhances international security.[285] The declaration recommended measures such as legal reviews to ensure that military AI capabilities are used consistent with States' international legal obligations, taking proactive steps to minimize unintended bias, and ensuring that such capabilities have explicit, well-defined uses.[286] While unenforceable from an international law perspective, such declarations form the core of a budding soft-law doctrine governing State use of AI technologies.

The need for international governance

The UN Secretary-General, the UN High Commissioner for Human Rights, and the newly established High-Level Advisory Body on Artificial Intelligence, have all stressed that leveraging the potential of AI to advance the rights and interests of all humanity[287] without discrimination requires a 'global, collaborative approach'[288] anchored in human rights. The Special Rapporteur on privacy has also suggested the creation of a specific ad hoc international law mechanism to examine transnational issues in ICT.[289]

By its very nature, AI poses trans-boundary risks. AI models, like all digital services, can easily cross borders and the risks associated with accidents or the harmful misuse of AI systems are not contained within one single jurisdiction. Attempts to regulate both the development and deployment of AI systems in one State therefore do no immunize it from harms that arise from AI-systems developed in states with less exacting regulations. Indeed, some entrepreneurs have even capitalized on this reality by proposing floating server farms in international waters that would be subject to no regulatory oversight of any kind.

Stakeholders operating in places with strong regulations have an especially pronounced interest that regulatory responses to AI be harmonized globally. Standardized regulations help level the playing field by promoting regulatory coherence. Such global standards can also provide greater legal certainty to businesses operating in different jurisdictions. A level playing field helps safeguard against a regulatory race to the bottom, in an attempt by certain jurisdictions to attract corporate investment at the cost of globally eroded product safety. Last but not least, companies also have an interest in securing consumer and user trust, as demonstrated by their own promotion of self-regulatory policies to ensure product safety. This is especially true for those corporations that continue to operate in regulated environments who have a strong interest in cultivating a positive reputation among consumers, not just for their particular corporate brand, but for AI as a whole, untainted by the impact of a few unscrupulous founders who consciously seek to evade regulatory efforts. Such global regulatory coherence can also help accelerate innovation by facilitating cross-border trade of AI-related products and services.

Game Theoretical principles reinforce the case for the international governance of AI. The well-known Prisoners' Dilemma problem illustrates the cumulative impact of individual incentives to defect from a "collaborative game." Just like an individual prisoner tempted to secure a plea deal for himself at the expense of his fellow prisoners, countries also might be tempted to enact a permissive regulatory regime within their borders to give their national AI industry a silent "boost" in light of other, more stringent regulatory regimes slowing the growth of potential competitors in other jurisdictions. Global regulatory measures, on the other hand, offer a way to make global collaboration more 'safe' for policy makers keen to encourage an innovation economy within their national borders.

Diplomacy tends to be a slow-moving train that is not always well-equipped to deal with the speed of technological innovation. Managing these geopolitical risks requires immediate and sustained collaboration, even despite our current-day diplomatic context marked by ideological polarization, mutual recrimina-

277. Conference Summary, Generative Artificial Intelligence Algorithm Regulation, National People's Congress Future Rule of Law Research Institute, April 2023, https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117#:~:text=policymakers%20are%20actively-,discussing,-how%20to%20maintain.

278. G7, "Charlevoix Common Vision for the Future of Artificial Intelligence," (2018), https://www.mofa.go.jp/files/000373837.pdf.

279. GPAI: The Global Partnership on Artificial Intelligence, Responsible AI, (accessed Dec. 28, 2023), https://gpai.ai/projects/responsible-ai/.

280. Lewin Schmitt, Mapping Global AI Governance: a Nascent Regime in a Fragmented Landscape, Vol.2 AI & Ethics 303-314 (2022), https://link.springer.com/article/10.1007/s43681-021-00083-y.

281. G20 Ministerial Statement on Trade and Digital Economy, (Jun. 2019), https://wp.oecd.ai/app/uploads/2021/06/G20-AI-Principles.pdf.

282. G20 New Delhi Leaders' Declaration, (Sep. 2023), https://www.consilium.europa.eu/media/66739/g20-new-delhi-leaders-declaration.pdf.

283. Executive Office of the President, "OMB Releases Implementation Guidance Following President Biden's Executive Order on Artificial Intelligence," (Nov. 1, 2023), https://www.whitehouse.gov/omb/briefing-room/2023/11/01/omb-releases-implementation-guidance-following-president-bidens-executive-order-on-artificial-intelligence/#:~:text=To%20ensure%20that%20agencies%20establish,and%20safety%20of%20the%20public.

284. Id.

285. Bureau of Arms Control, Deterrence, and Stability, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," (Nov. 9, 2023), https://www.state.gov/wp-content/uploads/2023/10/Latest-Version-Political-Declaration-on-Responsible-Military-Use-of-AI-and-Autonomy.pdf.

286. Id.

287. Office of the Secretary-General's Envoy on Technology, "High-Level Advisory Body on Artificial Intelligence", (accessed Dec. 28, 2023), https://www.un.org/techenvoy/ai-advisory-body.

288. OHCHR, "Türk calls for attentive governance of artificial intelligence risks, focusing on people's rights," (Nov. 30, 2023), https://www.ohchr.org/en/statements-and-speeches/2023/11/turk-calls-attentive-governance-artificial-intelligence-risks.

289. Joseph A. Cannataci, "Artificial intelligence and privacy, and children's privacy," (Jan. 25, 2021), Report of the Special Rapporteur on the right to privacy, Human Rights Council, 46th Session (22 Feb – 19 Mar, 2021; Agenda item 3), https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/015/65/PDF/G2101565.pdf?OpenElement.

tions, and strained trust. Luckily, policy makers in the United Nations and in other global forums are aware of the urgency and hard at work negotiating precisely such transnational regulatory regimes.

Ongoing debates and discussions demonstrate the difficulty of regulating AI. States, especially less developed states in the Global South, have a strong interest in pooling international research and expertise to identify best-practice responses to emerging AI technologies. At a minimum, there is a clear value in building consensus within the scientific community about the risks as well as the potential of AI technologies, as well as best-practices to steer the technology in a socially-beneficial direction. The challenge of creating a global regime to manage AI is reminiscent of the debates leading up to the creation of the International Panel on Climate Change (IPCC). That process began with a periodic series of high-level diplomatic meetings, fueled by global experts pooling up-to-date scientific research on climate change. Those annual conferences have since become the primary vehicle for efforts to accelerate and harmonize national climate change prevention and response policies. This approach, while slow, laborious, and arguably always falling just short of the "ideal" intervention, nonetheless has proven to be an effective tool to nudge policy-makers in a certain direction, primarily by opening avenues for cross-jurisdictional regulatory mimicking and facilitating normative cascades from one country to another. A similar effort might be warranted to shape a shared approach to AI and related technologies.

Drawing further on the IPCC process and the once-annual Conference of Parties Climate Change Summits, international governance efforts can also serve as a social 'multiplier of force,' convening not just scientists and policy makers, but also civil society activists, the media, celebrities, corporations, and even corporate lobbyists into one thematic conversation about the future. Such events can elevate policy discussions out of the rarified world of elite policymakers and technocrats into the popular consciousness, where attitudes also need to change in order for any reforms to be truly sustainable. Such an approach, could serve to promote greater digital literacy, incentivize efforts to share the benefits of technical and scientific knowledge globally, highlight examples of AI technology that improve the common good, and help equip the public with the requisite knowledge to demand that governments and private actors protect them from the potential harms of AI systems and algorithmic decision-making more broadly.

Finally, international governance can also serve a monitoring role, ensuring some level of state accountability for their use of AI as well as for the activities of private and non-state actors operating within their jurisdictions. Only global governance can lead to the creation of increasingly binding sources of international law that can help supplement the voluntary efforts at self-regulation driven by regulatory sharing of best practices. Such international normative developments will be key to harmonizing corporate responsibility in the field of AI, as argued by the OHCHR's B-Tech initiative.

The existing human rights corpus must continue to serve as the bedrock upon which all such international approaches are built in order to prevent societal harm and promote a greater and more equal enjoyment of technological progress by all.

The Role of the International Human Rights System in Global AI Governance

The value of the human rights-based approach to AI governance

During a high-level side-event of the 54th session of the Human Rights Council, the UN High Commissioner for Human Rights, Volker Türk, stressed that the question of which "limits should be [placed] on artificial intelligence and emerging technologies is one of the most pressing faced by society, government and the private sector." He urged for there to be a transition from a largely legalistic corporate risk-management compliance approach, "focusing largely on self-regulation and self-assessment by AI developers" towards a human rights-based approach that "embeds human rights in AI's entire lifecycle."[290] As Türk explains it, a human-rights based approach, which resembles the HRBA@Tech model described above, would embed human rights principles "[f]rom beginning to end" starting "in the collection and selection of data, as well as the design, development, deployment and use of the resulting models, tools and services."[291]

The High Commissioner concludes that "we need to resist the temptation to let the AI industry [. . . .] assert that self-regulation is sufficient, or to claim that it should be for them to define the applicable legal framework."[292] Turk claims to have learned his lessons from the tech industry's previous claims to be able to appropriately self-regulate in the domain of social media. "Whilst [corporations'] input is important, it is essential that the full democratic process – laws shaped by all stakeholders – is brought to bear, on an issue in which all people, everywhere, will be affected far into the future."[293]

To-date, efforts to ensure that AI remains aligned with human interests and values have largely relied on self-regulatory corporate AI principles, which rarely convincingly integrate human rights safeguards. Some, such as Google's or Salesforce's AI principles, explicitly commit to upholding the international human rights norms, while others' such as those of Microsoft or Meta specifically reference certain rights in the context of their AI activities.

These AI principles have proven to be more effective than early critics of corporate 'ethics-washing' believed. The Office of the High Commissioner for Human Rights, for example, has acknowledged that"despite early skepticism from external stakeholders about the sincerity of such high-level principles, many companies are successfully utilizing them to ground and guide AI product development and the remit of responsible AI teams whose role is to help the company implement its principles in practice."[294] Notwithstanding, there is still a case to be made that principles guiding responsible business conduct should be human rights principles and that the international human rights system and normative framework are suited as he normative anchor for any such self-governance efforts. Moreover, grounding corporate governance systems in the language of global human rights also serves as an insurance policy against diluted standards when market conditions turn more competitive, as arguably happened in 2023, when massive layoffs in the technology sector heavily eroded some of the big tech company's capacity (not to mention financial commitment) to address some of these knotty issues.

As others have argued, "there is no conflict between ethical values and human rights, but the latter represent a specific crystallization of these values that are circumscribed and contextualized by legal provi-

290. Volker Türk, "Artificial intelligence must be grounded in human rights, says High Commissioner," (Jul. 12, 2023), Comments at the High Level Side Event of the 53rd Session of the Human Rights Council, https://www.ohchr.org/en/statements/2023/07/artificial-intelligence-must-be-grounded-human-rights-says-high-commissioner.

291. Id.

292. Id.

293. Id.

294. United Nations Human Rights Office of the High Commissioner, "Responsible AI and Human Rights: An Overview of Company Practices Supplement to B-Tech's Foundational Paper on the Responsible Development and Deployment of Generative AI," (Oct. 24, 2023), https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/overview-human-rights-and-responsible-AI-company-practice.pdf.

sion and judicial decisions."[295] Indeed, human rights norms are universal, legally-binding and shared by peoples from all over the world. Of course there are important ongoing debates, especially in academic settings, about the extent to which human rights norms are truly universal in their origins, but by and large, the broadly acknowledged universal acceptance of the human rights discourse allows it to serve as the most authoritative set of values to harmonize AI governance approaches and facilitate international collaboration among various stakeholder groups. Ethical principles can be helpful to give a normative backdrop to legal and human rights, but not as replacements for those norms once their legitimacy has already been established. In other words, 'don't fix what's not broken.'

The human rights normative framework has also proven remarkably able to address unforeseen challenges. It has grown from its much narrower state-centric origins in the 1940s to become a multifaceted corpus of laws, rules, and standards thqt have increasingly become the domain not just of states, but also civil society groups, international organizations, academics and corporations. Part of this success stems from an approach that balances individual rights with competing legitimate social or national interests.[296] This same approach will be important in mitigating the risks associated with AI-systems without unduly hampering innovation in an area with so many potential upsides. The human rights corpus itself is likely to evolve over time in light of the challenges posed by AI and other new and emerging technologies – all while remaining tethered to the original values of the Universal Declaration of Human Rights and other foundational norms that gave birth to the human rights movement.

A holistic approach to AI that embraces both the negative as well as the positive implications on human rights protections of this new technology allows ample space for discussions of how to make the world a better place. It allows human rights practitioners not only to brace themselves for the potential negative impacts of AI, but also to embrace AI as a tool for social progress. In many quarters, human rights in the domain of new and emerging technologies are associated with actions that the state (or corporations) should not do (for example to violate labor rights or infringe citizens' rights to privacy). Economic, social and cultural (ESC) human rights, however, also include numerous rights that states are obligated to progressively realize. AI can support the pursuit of such ESC human rights (such as the right to education, shelter, food and spreading the fruits of scientific discovery).[297] Indeed, AI can help advance such rights in ways previously unimaginable by humans relying only on their own capacities.

Finally, the past efforts of human rights practitioners have led to the development of an increasingly robust ecosystem of institutions and formal, informal, and norms-based processes that collectively work towards the implementation of human rights standards and accountability and redress for victims of human rights abuse. These systems are constantly evolving and (hopefully) improving. That said, they already provide a strong basis for businesses and States working with AI to orient their conduct, for example by conducting rights-centric impact assessments of new and emerging technologies, engaging in meaningful consultations with vulnerable communities, or designing effective grievance processes. These actions, guided by existing human rights standards, apply equally to the technology sector as they do to more traditional businesses.

The rapidly evolving AI policy landscape at the United Nations

In the words of UN Secretary-General António Guterres, there is a "wide and growing" gap between AI developments and our collective governance capacities that require the global community to "play[] catch up" to get back "ahead of the wave."[298] In recent years, the Secretary-General has consistently focused attention on the need for coordinated action on AI. In his 2020 Road map for digital cooperation, Guterres underlined the existence of multiple frameworks (over 160 sets of AI ethics and governance principles worldwide) and the absence of a common platform.[299] Guterres highlighted the need for new solutions, not new principles. He suggested that the principles for governing AI should be based on existing obligations under the UN Charter and the Universal Declaration of Human Rights.

To realize that vision of developing global oversight and addressing governance gaps, the UN Secretary General in November of 2023 launched a High-Level Multi-stakeholder Advisory Body on Artificial Intelligence. The Advisory Board's objective is to link various pre-existing AI governance initiatives, build a shared understanding of AI's risks and potential benefits, and leverage the technology's potential as a force for good and sustainable development.[300] The Advisory Board is mandated to consult with various stakeholders and build momentum towards a 'Global Digital Compact' that will ideally be finalized at the 'Summit of the Future' held in September of 2024. High on the agenda at this Summit will be the governance of AI and the development of the 'Global Digital Compact' as a central framework to align various national and regional governance approaches to AI.

The High-Level Advisory Board is also expected to deliberate about the potential for an international body tasked with the oversight of developments in the AI sector, drawing on the precedent institutions tasked with overseeing atomic energy (the International Atomic Energy Agency - IAEA), civil aviation (the International Civil Aviation Organization - ICAO), public health (the World Health Organization - WHO) or climate change (the International Panel on Climate Change - IPCC). Any initiative to create such an organization, which has been proposed by numerous senior AI executives and backed by the Secretary-General,[301] would require UN Member States to agree on the mandate and functions of such an agency. The Summit of the Future is intended as a venue for global leaders to recommit themselves to multilateral cooperation in the pursuit of a transnational approach to the development of trustworthy AI systems.

Even though currently the spotlight is on the High-Level Advisory Board and any proposals it may generate in advance of the 2024 'Summit of the Future,' a range of parallel efforts continue across the United Nations system.

One of the major initiatives is taking place at the International Telecommunication Union (ITU), which was founded in 1865 to help facilitate and coordinate communication globally. The ITU's 'AI for Good' platform, launched in 2017 in partnership with forty other UN agencies, convenes an annual summit to highlight potential AI applications that enable progress towards the achievement of the SDGs and to engage with the main stakeholders on how to harness the potential of those solutions. The platform, which brings together industry representatives and policy-makers, will hold an 'AI Governance Day' ahead of its next summit scheduled in May 2024, where it will explore various safeguards for the responsible development of AI.[302] Besides acting as a convener of global expertise on AI, the ITU's standard-setting work has also addressed the impacts of AI on areas such as health[303] environmental efficiency.[304]

The Organization for Economic Cooperation and Development (OECD), which is not part of the United Nations but supports the UN's development mission, in 2019 launched the first intergovernmental standards on AI, to ensure that AI benefits 'society as a whole.' The Recommendation on Artificial Intelligence (OECD AI Recommendations) are intended to promote inclusive and trustworthy AI that upholds human rights and democracy.[305] It sets forth five values-based "Principles for the Responsible Stewardship of Trustworthy AI", along with five corresponding recommendations for policy makers. The 38

295. Mantelero, A., Esposito, S., An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems, Computer & Security Law Review, 6 (2021), https://ssrn.com/abstract=3829759.

296. Kate Jones, "AI governance and human rights: Resetting the relationship," (Jan. 2023), International Law Programme, https://www.chathamhouse.org/2023/01/ai-governance-and-human-rights/03-governing-ai-why-human-rights.

297. Id.

298. United Nations, Secretary-General's Statement at the UK AI Safety Summit, (Nov. 2, 2023), https://www.un.org/sg/en/content/sg/statement/2023-11-02/secretary-generals-statement-the-uk-ai-safety-summit.

299. United Nations, Road map for digital cooperation: implementation of the recommendations of the High-level Panel on Digital Cooperation, (May 2020), https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/102/51/PDF/N2010251.pdf.

300. United Nations, Secretary-General Press Remarks launching High-Level Advisory Board on Artificial Intelligence, (Oct. 26, 2023), https://www.un.org/sites/un2.un.org/files/sgs_remarks_announcing_high-level_advisory_body_artificial_intelligence_26_october_2023.pdf.

301. Michelle Nichols, "UN chief backs idea of global AI watchdog like nuclear agency," (Jun. 12, 2023), Reuters,https://www.reuters.com/technology/un-chief-backs-idea-global-ai-watchdog-like-nuclear-agency-2023-06-12/.

302. Seizo Onoe, "How ITU Powers AI Action for Good," (Nov. 29, 2023), ITU, https://www.itu.int/hub/2023/11/how-itu-powers-ai-action-for-good/.

303. ITU, "Focus Group on 'Artificial Intelligence for Health'," (accessed Dec. 28, 2023), https://www.itu.int/en/ITU-T/focusgroups/ai4h.

304. ITU, "Focus Group on Environmental Efficiency for Artificial Intelligence and other Emerging Technologies (FG-AI4EE)," (accessed Dec. 28, 2023), https://www.itu.int/en/ITU-T/focusgroups/ai4ee/Pages/default.aspx.

305. OECD AI Principles overview https://oecd.ai/en/ai-principles.

OECD member states can be said to have endorsed the OECD AI Recommendations as non-binding soft-law standards. In June of 2019 the G20, which notably includes Russia, China, India, Indonesia, and Saudi Arabia, also endorsed the OECD AI Recommendations.[306] Numerous other countries, primarily in South America and Eastern Europe, have also pledged their adherence to the Recommendations. The OECD AI Recommendations thus serve as an emerging global standard for the development of trustworthy AI.[307] Mere commitments to such global standards, however, still mask a diversity of policy approaches to AI governance, and the implementation of concrete national policy measures to give effect to the OECD principles remains a work in progress.[308]

In 2020, the UN's High-level Committee on Programmes, which oversees coordination and policy coherence across the UN system, established the Inter-Agency Working Group on Artificial Intelligence (IAWG-AI). The IAWG-AI has different workstreams focusing on topics such as human rights, education, justice, and capacity-building. It has developed the Principles for the Ethical Use of Artificial Intelligence in the United Nations System[309] which, although intended to guide the design, development, deployment, and use of AI within the UN system alone, also provides an example of an ethical approach to AI that can orient the development of other governance frameworks globally.

The first truly global effort to develop AI ethics standards was developed under the auspices of the UN Educational, Scientific and Cultural Organization (UNESCO) in November 2021 and subsequently adopted by UNESCO's 193 Member States. The UNESCO Recommendation on the Ethics of Artificial Intelligence[310] develops a human rights-based approach to the ethics of AI based on four 'values, defined as "motivating ideals" that "shap[e] policy measures and legal norms," and ten corresponding "principles" that "unpack" those core values in terms that can be more easily operationalized by policy makers.[311] The UNESCO Recommendations place the dignity and non-objectification of humans at the center of their approach to AI systems. They call for the development of legislation anchored in existing human rights obligations, multilateral cooperation to address potential harms, consultation and joint capacity-building across national borders, and the development of norms-based approaches to AI governance such as the development of global AI certification mechanisms.[312]

Other efforts to influence emerging AI governance frameworks include the work of the Policy Network on Artificial Intelligence (PNAI), hosted by the UN's Internet Governance Forum. The PNAI's focus is on capturing the benefits of AI for the Global South. PNAI's initial report, published in October 2023,[313] is expected to contribute to the discussions around the Global Digital Compact. Its initial recommendations include the development of standards that center around human dignity, human rights and gender equality, justice, well-being, diversity, social and economic development, and sustainability. PNAI has also called for increased participation by stakeholders from the Global South in discussions about AI governance.

306. G20 AI Principles, (Jun. 29, 2019), https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf.

307. OECD, Recommendation of the Council on Artificial Intelligence, (Nov. 8, 2023), OECD/LEGAL/0449, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

308. OECD, "The state of implementation of the OECD AI Principles four years on," (2023), OECD Artificial Intelligence Papers, No. 3, OECD Publishing, Paris, https://doi.org/10.1787/835641c9-en.

309. United Nations, Principles for the Ethical Use of Artificial Intelligence in the United Nations System, (2022), High-Level Committee on Programmes (HLCP) Inter-Agency Working Group on Artificial Intelligence, https://unsceb.org/principles-ethical-use-artificial-intelligence-united-nations-system.

310. UNESCO, Recommendation on the Ethics of Artificial Intelligence, (Nov. 23, 2021), https://unesdoc.unesco.org/ark:/48223/pf0000381137.

311. Id., at par. 10. The 4 Values are: (1) the Respect, protection and promotion of human rights and fundamental freedoms and human dignity, (2) the Environment and ecosystem flourishing, (3) Ensuring diversity and inclusiveness, and (4) Living in peaceful, just and interconnected societies. The corresponding Principles are (1), Proportionality and Do No Harm, (2) Safety and Security, (3) Fairness and Non-Discrimination, (4) Sustainability, (5) Right to Privacy and Data Protection, (6) Human Oversight and Determination, (7) Transparency and Explainability, (8) Responsibility and Accountability, (9) Awareness and literacy, and (10) Multi-stakeholder and adaptive governance and collaboration.

312. Id. at par. 56.

313. Internet Governance Forum Policy Network on Artificial Intelligence, "Strengthening multi-stakeholder approach to global AI governance, protecting the environment and human rights in the era of generative AI," (Oct. 2023), https://www.intgovforum.org/en/filedepot_download/282/26545.

Current Guidance from the Human Rights System on AI Governance

Human Rights Council resolutions on new and emerging technologies

In recent years, the Human Rights Council (HRC), led by the "Core Group" of the Republic of Korea, Austria, Brazil, Denmark, Morocco, and Singapore, has increasingly focused its efforts on evaluating the human rights implications of new and emerging technologies, including AI.

In 2019, the Council adopted resolution 41/11,[314] in which it recognized that digital technologies have the potential to accelerate human progress and to facilitate efforts to promote and protect human rights. The resolution noted that the possible human rights impacts of these technologies are still poorly understood, and requested the HRC's Advisory Committee to prepare a report exploring the human rights implications of new and emerging digital technologies as well as the potential role of international human rights mechanisms in helping to address those issues.

The Advisory Committee presented its subsequent report to the HRC in June 2021.[315] The report defines "new technologies" as the technological innovations that transform the boundaries between the virtual, physical, and biological spaces, and included in that definition new technologies and techniques for the datafication (the process of transforming subjects, objects, and practices into digital data), data distribution, and automated decision-making. Examples of such technologies include AI, the Internet of Things, blockchain technology, and cloud computing, amongst others. The Advisory Committee report noted the paradox of new and emerging technologies (see above), as well as the seminal role of private actors in the development of these technologies. The report highlighted various gaps in the existing human rights framework's ability to address the impacts of new and emerging digital technologies according to a unified and globally accepted approach. These included unresolved philosophical questions, a lack of cooperation and contact between the human rights and tech communities, and a selective preoccupation on only a subset of technologies and human rights harms, as opposed to a more holistic approach embracing also the potential 'upsides' of technological innovation. The report also noted the practical challenges of responding to the breakneck pace of scientific innovation and the fragmentation of regulatory initiatives across the global governance landscape. The report concluded by calling for a nuanced outcome-focused "human rights-based approach" emphasizing a holistic understanding of technology, an equally holistic approach to human rights, and finally a commitment to holistic governance and regulatory efforts.

Previous efforts to deal with human rights and ICT technologies have often unfolded in disciplinary silos. The Advisory Committee report emphasizes the importance of more cross-disciplinary solutions. It also rejects the premise that technology is inherently value-neutral. This suggests that efforts should be made to eliminate structural bias from new and emerging technologies. The same insight also opens the possibility to design new and emerging technologies to be not just value neutral (ensuring that they 'do no harm'), but rather to to be intentionally biased in favor of human rights outcomes (to 'make the world a better place'). The Advisory Committee Report also recognizes the futility of focusing on the human rights impacts of any one technology in isolation, and insists that a better approach would evaluate an interconnected ecosystem of technologies. This would represent a departure from previous approaches that tended to focus on only one technology in isolation.

Building on the Advisory Committee's report, the HRC issued Resolution 47/23, in which it reaffirmed the need for a human rights-based approach to new and emerging digital technologies. The resolution further requested that OHCHR to convene expert consultations exploring the links between the human rights

314. Human Rights Council Resolution 41/11, "New and emerging digital technologies and human rights," (Jul. 17, 2019), UN Doc. No. A/HRC/RES/41/11, https://undocs.org/Home/Mobile?FinalSymbol=A%2FHRC%2FRES%2F41%2F11&Language=E&DeviceType=Desktop&LangRequested=False.

315. Human Rights Council Advisory Committee, "Possible impacts, opportunities and challenges of new and emerging digital technologies with regard to the promotion and protection of human rights," (June 2021), UN Doc. No. A/HRC/47/52, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/110/34/PDF/G2111034.pdf.

impacts of new and emerging technologies and technical standard-setting.

In June 2023, the 'Core Group' narrowed its focus on AI systems. The HRC's Resolution 53/29, which was adopted by consensus, builds on the HRC's balanced and outcome-driven approach to NETs, highlighting AI's potential to both threaten but also promote and protect human rights. The resolution focuses on AI's potential to "facilitat[e] access to information and participation in public life, strengthen[] the efficiency and accessibility of health-care services, enable[] greater availability and accessibility of education, advance[e] gender equality and empower[] all women and girls, contribut[e] to the full enjoyment of human rights by older persons, persons with disabilities and those in vulnerable situations, strengthen[] climate mitigation and adaptation and support[] environmental protection", while also recognizing that "certain application of AI present an unacceptable risk to human rights."[316]

In line with proposals made in the context of the HRBA@Tech model, the resolution emphasizes the importance of a human rights-based approach to new and emerging digital technologies[317] by protecting individuals from harm, notably through human rights due diligence and impact assessments, guarding against discrimination and bias, promoting algorithmic transparency, ensuring that data collection, storage and use is consistent with human rights obligations, and strengthening oversight and enforcement capacity. The resolution also encourages multi-stakeholder collaboration and a further exploration of ways for the Human Rights Council to promote the human rights based approach to AI. The resolution requested for the OHCHR to undertake a gaps-analysis of the various efforts underway at the UN to grapple with the human rights implications of AI and to build its own capacity to support national and private corporate efforts to promote trustworthy AI.

Guidance from the UN Human Rights System

The UN High Commissioner for Human Rights Volker Türk has described the rapid advances in AI as a "paradox of progress,"[318] holding the potential to transform lives and help solve complex challenges while also potentially undermining human rights and human dignity. Previous explorations of AI's impact by various actors in the human rights system have tended to be fragmented; largely driven by the mandates of the particular institution or agency driving the discussion. Given most human rights institutions' mandates to raise the alarm about emerging human rights threats, most of the work on AI so far has naturally tended to focus on the negative human rights impacts of AI and other new and emerging digital technologies.

Despite this predominant focus on the human rights downsides of new and emerging technologies, there is also at least a rhetorical consensus that AI technologies can also be used to promote and protect human rights. Various reports have alluded to this potential, even if only in passing. The Special Rapporteur on persons with disabilities, for example, has highlighted the potential of AI systems to improve accessibility through assistive and mobility-enhancing technologies.[319] He further pointed to with potential of other AI-assisted technologies, such as adaptive learning platforms, one-to-one tutoring, and speech recognition applications to enable persons with disabilities to interact socially and professionally with others and access information and education opportunities that previously were not available to them.[320] Similarly, the Independent Expert on the Rights of Older Persons noted AI's ability to help older persons live autonomously.[321] Supported decision making devices, for example, can help an older individual review options for daily life choices and make decisions accordingly, while self-learning technology may learn what an older person or someone with communications difficulties wishes, and assist them in communicating their needs. The UN Special Rapporteur on freedom of expression has also examined the beneficial impacts of AI on communication. He pointed at AI's capacity to provide improved and personalized access to information and services, increasing users' ability to access information in several languages and to be exposed to content that is relevant to their experiences and preferences.[322] The Human Rights Council Advisory Committee also emphasized the potential human rights benefits of AI-enhanced technologies, for example their potential to improve educational opportunities for deaf children, and level linguistic, geographic, cultural and societal barriers (for example by enabling persons living alone or in remote areas to interact with others through remote telepresence and companion robots, or by increasing women's access to education opportunities through online solutions).[323] Various other commentators have pointed out the potential for AI-systems to improve service delivery by virtue of their ability to optimize existing government and social welfare processes. For example, former Special Rapporteur on extreme poverty, Philip Alston stressed the potential of digital technologies, including artificial intelligence, to "improv[e] the well-being of the less well-off members of society," albeit but not without "deep changes in existing policies [towards a more] genuine commitment to [. . . .] ensure a decent standard of living for everyone in society."[324]

By far the more predominant focus of the human rights community has been on the human rights downsides of new and emerging technologies, including AI. For example, the Special Rapporteur on racism, the CERD and the Special Rapporteur on persons with disabilities have identified numerous instances where the use of algorithmic systems have led to discriminatory outcomes, including in access to health care, justice, or employment opportunities. Likewise, the Special Rapporteur on the rights of persons with disabilities has considered the potential for AI-powered recruitment systems to discriminate against persons with disabilities or with special needs. This may happen, for instance, when such systems fail to consider reasonable accommodations and the need for assistive technologies of candidates with disabilities, or when automated interviewing systems misread facial and verbal expressions of persons with disabilities and special communications needs.[325]

Several Human Rights Mechanisms have highlighted particularly risky application of AI-systems. Examples include AI-powered facial recognition technologies used by law enforcement agencies, the deployment of "predictive policing" strategies to prevent crime, and censorious content moderation practices.[326] Many AI systems are based on inherently biased datasets. While they are thus described by their proponents as being "data-driven" and "objective," in fact they also might amplify historically biased correlations between certain races, genders, religions or other protected categories with perceived predilections towards criminal behavior, thus focusing the limited resources of law enforcement agencies and courts even more on these communities. The UN Special Rapporteur on Racism noted how the inherent opacity of these systems only exacerbates these tendencies.[327]

Linked to concerns about the discriminatory impacts of AI technologies, several mechanisms have also noted risks to privacy rights of AI systems, particularly in the provision of healthcare and social welfare services. The former Special Rapporteur on extreme poverty has flagged the "dystopian" specture of "welfare states turn[ing] into digital welfare states," where there is an overall lack of transparency surrounding the systems' data collection methods, the use of the data, and the technologies being deployed.[328]

316. HRC, supra note 1, preamble.

317. Id., at Art. 3.

318. Volker Türk, "Türk calls for attentive governance of artificial intelligence risks, focusing on people's rights," (Nov. 30, 2023), Speech given at the Generative Artificial Intelligence and Human Rights Summit, https://www.ohchr.org/en/statements-and-speeches/2023/11/turk-calls-attentive-governance-artificial-intelligence-risks.

319. Gerard Quinn, Report of the Special Rapporteur on the rights of persons with disabilities, Human Rights Council, UN Doc. No. A/HRC/49/52, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/397/00/PDF/G2139700.pdf.

320. Id.

321. Rosa Kornfeld-Matte, Report of the Independent Expert on the enjoyment of all human rights by older persons, Human Rights Council, UN Doc. No. A/HRC/36/48, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G17/219/52/PDF/G1721952.pdf.

322. David Kaye, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, UN Doc. No. A/73/348, https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf.

323. Human Rights Council Advisory Committee, supra note 316.

324. Philip Alston, Report of the Special rapporteur on extreme poverty and human rights, Human Rights Council, UN Doc. No. A/74/48037, https://digitallibrary.un.org/record/3834146/files/A_74_493-EN.pdf?ln=en.

325. Gerard Quinn, supra note 320; Committee on Elimination of Racial Discrimination, "General Recommendation No. 35. Preventing and Combating Racial Profiling by Law Enforcement Officials," (Dec. 17, 2020), https://documents-dds-ny.un.org/doc/UNDOC/GEN/G20/349/50/PDF/G2034950.pdf.

326. E. Tendayi Achiume, "Racial and xenophobic discrimination and the use of digital technologies in border and immigration enforcement," (2021), Human Rights Council, A/HRC/48/76, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/379/61/PDF/G2137961.pdf?OpenElement; David Kaye, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, (2018), Human Rights Council, UN Doc. No. A/73/348, https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf.

327. Achiume, supra note 327.

328. Alston, supra note 325.

Several human rights mechanisms have provided guidance to stakeholders on policies and practices that serve to promote and protect human rights. Most of these recommendations have been addressed to States and have revolved around calls to ensure that regulations are grounded in legally binding international human rights principles and standards.

Additional guidance has focused on States' obligations to regulate the use and sale of AI systems that present clear human rights risks. Former UN High Commissioner for Human Rights, Ms. Michelle Bachelet, noted in 2021 that "the higher the risk for human rights, the stricter the legal requirements for the use of AI technology should be."[329] Other Mechanisms have recommended imposing moratoria or prohibitions on the use and sale of AI tools when these pose significant human rights risks, at least until those risks can be assessed and mitigated.[330] This argument has been made, for instance, with regard to real-time facial recognition technologies, which have raised privacy rights concerns and heightened concerns about their discriminatory potential. Similarly, in a 2021 press release, the High Commissioner has called for the banning or suspension of AI technologies until there are sufficient safeguards in place for them to operate in compliance with international human rights law.[331]

Other more specific guidance has included calls for AI-systems used in law enforcement to be designed transparently, while providing access to researchers and civil society actors so that they may assess the source code.[332] The Committee on the Elimination of Racial Discrimination has also recommended the establishment of representative oversight mechanisms to ensure that Government AI systems are human rights compliant.[333] The Office of the High Commissioner for Human Rights has recommended that States develop regulations for the use of AI-systems commensurate to the human rights risks, focusing notably on law enforcement, national security, criminal justice, social protection, employment, health care, education, and the financial sector as key priority areas.[334]

Mechanisms have also recalled the State's duty to protect from harm arising from business practices. The Special Rapporteur on freedom of expression, for example, has encouraged States to ensure that laws and policies related to AI not focus solely on public sector regulation but also on private sector AI applications.[335] Similarly, the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance has called on States to "ensure that ethical frameworks and guidelines developed to provide flexible, practical and effective regulation and governance of emerging digital technologies are grounded in legally binding international human rights principles."[336] Similarly, former High Commissioner Michelle Bachelet has noted that the State's role in shaping the development and use of AI should go beyond a legal and policy-setting function, and include working with private sector AI developers and service providers to ensure AI's compliance with human rights obligations.[337]

The UN Guiding Principles (UNGPs) are typically understood as the dominant framework guiding the role of the private sector as an integral part of the human rights system. Interventions have placed particular emphasis on the role of due diligence to identify potential human rights risks, along with the related duty to take preventative and mitigating action to avoid adverse human rights impacts. Several Special Procedure mandate holders have called for their focus area to be covered in any standard due diligence process. The Special Rapporteur on the rights of persons with disabilities, for example, has called for any due diligence process to be "inclusive of disability."[338]

Several calls have also been made to improve remedies for individuals and communities harmed by AI. The Special Rapporteur on the rights of persons with disabilities has recommended businesses to 'ensure accessible and effective non-judicial remedies and redress for human rights harms arising from the adverse impacts of artificial intelligence systems on persons with disabilities.' Similarly, the High Commissioner, examining the right to privacy in the digital age, has called on businesses to provide remedy and cooperate in remediation processes in cases they have caused or contributed to adverse human rights impacts, as well as to establish internal grievance redress mechanisms.[339]

The Working Group on Business and Human Rights has provided extensive guidance on how businesses can implement the UNGPs. There remains some need to clarify how this guidance applies to tech companies, however. In this regard, in May 2023, High Commissioner Türk has clarified that human rights due diligence must apply to the design, development and use of technology products and services.[340]

The B-Tech project

The OHCHR in 2019 launched its B-Tech project in order to provide more specific guidance about the application of the UNGPs to tech companies.[341] The project's objective was to identify and mitigate the risks of new and emerging digital technologies, and to harness their potential as a force for good. The project is intended to provide guidance on four strategic themes[342]:

1. Addressing human rights risks in business models;
2. Human rights due diligence and end-use;
3. Accountability and remedy; and
4. An exploration of "a Smart Mix" regulatory and policy responses to human rights challenges linked to digital technologies.

Since its inception, B-Tech has engaged with business leaders, investors, civil society, and governments. It has facilitated numerous multi-stakeholder discussions on the implementation of the UNGPs in the technology sector, including from regional and sectoral perspectives.

The B-Tech Project has generated significant guidance for the tech sector, ranging from more general descriptions of business responsibilities to specific guidance on particular issues of concern. The Project has underscored the primary responsibility of the State as the main duty bearer of human rights obligations, as set out in international law and in the UNGPs. States are responsible for the regulation of technology companies to prevent them from violating human rights, and of course also obligated to themselves refrain from rolling back applicable human rights protections. States have a duty to protect the right to privacy that cannot be undermined by technological developments such as AI-driven surveillance tools. B-Tech has also noted the State's duty to ensure that any companies with which it works, or which receive State subsidies, abide by their human rights obligations.[343]

One of the core themes of the B-Tech Project is the exploration of a 'smart mix of measures' combining international and domestic regulatory and policy approaches to the development and deployment of trustworthy AI. This approach recommends that any policy contain both 'carrots' and 'sticks' to incentivize corporate best practices. B-Tech has produced extensive guidance on human rights due diligence, stressing that these processes should not be seen as "simplistic compliance exercise[s]" but rather as a complex practice that can question a company's entire business model, push it to pursue different approaches to its operations, and to continually improve its product

329. UN News (author unknown), "Urgent action needed over artificial intelligence risks to human rights," (Sep. 15, 2021), https://news.un.org/en/story/2021/09/1099972.

330. Quinn, supra note 320.

331. Office of the High Commissioner for Human Rights, Press Release (Sep. 15,2021), https://www.ohchr.org/en/2021/09/artificial-intelligence-risks-privacy-demand-urgent-action-bachelet.

332. Committee on Elimination of Racial Discrimination, supra note 326.

333. Id.

334. OHCHR, "The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights," (2021), A/HRC/48/31, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/249/21/PDF/G2124921.pdf.

335. Kaye, supra note 327.

336. Human Rights Council, Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance, (Jun. 18, 2020), 44th Session (15 June – 3 July, 2020, Agenda Item 9), A/HRC/44/57, 58., https://digitallibrary.un.org/record/3883249?ln=en.

337. OHCHR, supra note 335, at par. 51.

338. Quinn, supra note 320.

339. Id.

340. Volker Türk, "Global Digital Compact must be guided by human rights," (May 8, 2023), https://www.ohchr.org/en/statements/2023/05/global-digital-compact-must-be-guided-human-rights-says-turk.

341. Office of the High Commissioner for Human Rights, "B-Tech Project," (accessed Dec. 28, 2023) https://www.ohchr.org/en/business-and-human-rights/b-tech-project.

342. Id.

343. Office of the High Commissioner for Human Rights, "Bridging Governance Gaps in the Age of Technology – Key Characteristics of the State Duty to Protect a B-Tech Foundational Paper," par. 1-5 (2021), https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/b-tech-foundational-paper-state-duty-to-protect.pdf.

design, including through collaboration with a wide range of stakeholders.[344]

The B-Tech project has also paid considerable considerable attention to the development of company-based grievance redress mechanisms as a means for individuals and communities to seek accountability in response to large-scale human rights violations.[345] Grievance redress mechanisms established proactively by companies also function as early warning systems; for instance when complaints received through dedicated grievance mechanisms or hotlines begin to show systemic human rights violations associated with a particular technology. The B-Tech project notes that the establishment of such mechanisms serves the corporate 'bottom line' by improving companies' reputations and avoiding the potentially significant financial and legal repercussions of human rights violations when they do occur.

The B-Tech project has already issued several recommendations based on particular case studies. It has intervened, for example, regarding the implementation and enforcement of the EU Digital Services Act, focusing on improving risk assessment, transparency and stakeholder engagement.[346] It has also developed specific regional guidance, such as the B-Tech Africa project.[347] Addressed specifically at the African tech sector, that project is focused specifically on the realization of the Sustainable Development Agenda in Africa and the digitalization of African markets.[348]

B-Tech has also launched a project to 'guide more effective understanding, mitigations and governance of the risks of generative artificial intelligence.' B-Tech's publications in this area include a taxonomy of human rights harms associated with generative AI. It has also begun to map regulatory responses and corporate best practice with regards to 'responsible AI.' Its work has resulted in the publication of a holistic approach to the development of trustworthy generative AI. Similarly to the above-mentioned HRBA@Tech approach put forward by this paper series, B-Tech's foundational paper also takes a multi-layered governance approach that focuses on different stakeholders' responsibilities in upholding a human rights-based approach to generative AI. The B-Tech paper also embraces the concept of a technology lifecycle, focusing on specific business processes, such as human rights impact assessments, algorithmic audits, or data quality reviews, that will be more effective at different points along that technology lifecycle. These are welcome developments that provide increasingly specific answers to the calls for more authoritative guidance on how to ensure that AI remain responsive to human rights concerns.

Next steps for the Human Rights Council

The international human rights system has broadly shown itself to be alert to the potential human rights implications of AI-systems, and has begun to generate increasingly concrete guidance to a range of stakeholders interested in the development and deployment of trustworthy AI. The Human Rights Council issued a number of important resolutions focusing on AI and human rights, highlighting the ongoing global appetite for multilateral cooperation in the development of a human rights-based approach to AI governance. OHCHR and the international human rights mechanisms have likewise made regular interventions on the human rights implications of AI. The current and previous High Commissioners for human rights, Volker Türk and Michelle Bachelet (respectively), have been vocal proponents of these various initiatives, while also recognizing the potential of AI to drive increased social well-being. Similarly, various Special Procedures have focused on how AI is affecting – or may soon affect – the rights covered by their mandates, while the Treaty Bodies have developed an increasingly broad corpus of work exploring how AI might impact rights discussed in their respective founding conventions.

According to the authors of this report, there are several traditional functions of the international human rights system:

- **The normative function** (the generation of guidance on the application of international human rights norms to new issues, and when necessary defining and building consensus around new norms);
- **The convening function** (bringing together various stakeholders to empower rights-holders, share information, and promote good practices and mutual understanding
- **The monitoring function** (monitoring compliance of State and non-State actors with human rights norms and alerting the parties involved as well as the international community whenever there are instances of misconduct;
- **The accountability function** (building out the adjudicative capacity of international institutions and fora to raise the cost of violations and identify avenues for accountability and redress;
- **The assistance function** (providing technical assistance and capacity-building to duty-bearers in order to strengthen their ability to live up to their human rights obligations and commitments);
- **The educational function** (informing and educating stakeholders of their rights and responsibilities); and
- **The mainstreaming function** (ensuring that human rights are properly integrated into global governance systems).

The response of the human rights system to AI, so far at least, has prioritized only a limited number of these functions. Speaking in generalities, the response so far has focused primarily on the convening, monitoring, and educational functions of the international human rights system. It is also making slow but incremental progress with regard to the normative function. By contrast, it has not yet made much progress with regard to the accountability, assistance, or mainstreaming functions. This rudimentary gaps analysis should guide future efforts to strengthen the existing human rights mechanisms with regard to AI without duplicating or replicating existing efforts.

Staying ahead of the curve

The HRC and the wider UN human rights system have been 'ahead of the curve' in terms of analyzing and drawing attention to the human rights risks associated with AI. The human rights system, so far at least, has been somewhat less effective at helping stakeholders design and deploy AI in ways that actively promote human rights, and it is only just beginning to embrace a multi-stakeholder culture of solution-oriented problem solving.

For the HRC and the broader human rights system to stay 'ahead of the curve,' it must now turn from the 'theoretical' to the 'operational,' and away from an exclusive focus on problem identification towards collaborative problem solving. As discussed at the 9th Glion Human Rights Dialogue, this should involve a number of steps clustered around two main priorities. Specifically the human rights community should:

1. Further clarify human rights norms as they relate to digital technology, including AI, and (critically) distill them into a single and easily-accessible normative framework for the benefit of States and technology companies (including SMEs);
2. Promote the uptake and implementation of those norms by States and technology companies by:
    a. leveraging the convening power of the HRC;
    b. promoting cooperation and dialogue at the international and national levels; and
    c. providing technical and capacity building assistance at the national-level.

In the opinion of the authors, the easiest and most obvious way to achieve these objectives is by the establishment of an additional thematic Special Procedure under the auspices of the Human Rights Council. There exist various different types of thematic Special Procedure mandates, including Special Rapporteurs, Independent Experts, and Working Groups. Speaking in generalities, these Special Procedures play three principal roles in the international human rights ecosystem.

1. Special Procedures can help to establish human rights norms as they relate to a specific issue or population group.

    States acting on their own are often incapable of effective policy making in response to emerging human rights concerns (often because of the politically divisive nature of these problems). It has therefore proven helpful in the past for States to 'outsource' the development of such normative regimes to independent experts mandated by the HRC to do so. Special Proce-

344. Office of the High Commissioner for Human Rights, "Key Characteristics of Business Respect for Human Rights: A B-Tech Foundational Paper," (2020), https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/key-characteristics-business-respect.pdf.

345. Office of the High Commissioner for Human Rights, "Designing and implementing effective company-based grievance mechanisms A B-Tech Foundational Paper," (Jan. 2021), https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/access-to-remedy-company-based-grievance-mechanisms.pdf.

346. B-Tech, CDT Europe, "Fostering responsible business conduct in the tech sector – the need for aligning risk assessment, transparency and stakeholder engagement provisions under the EU Digital Services Act with the UNGPs," (Aug. 2023), https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/B-Tech_Blog_RBC_DSA_UNGPs-alignement.pdf.

347. B-Tech Africa Project - A part of UN Human Rights B-Tech Project, (Sep. 2022), https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/2022-08-26/B-Tech-Africa-Project-one-pager.pdf.

348. See 'Events, meetings, etc.' section: https://www.ohchr.org/en/business-and-human-rights/b-tech-project.

dures serving this function have typically done so through the publication of annual reports to the Council, which are then deliberated upon and eventually – pending sufficient consensus – endorsed by the entire HRC. The development of the UNGPs stand as a prominent example of this process and a potential methodological precedent for the development of a future human rights based framework to govern new and emerging technologies.

2. Special Procedures are also well-placed to work with States, through cooperation and dialogue at the UN-level, but also at the national level (through country missions). In so doing, they can promote, support and secure the implementation of emerging human rights norms by national governments as well as by relevant non-State actors.

   The Human Rights Council typically mandates Special Rapporteurs with this task. In some cases it has also created five-person cross-regional Working Groups to fulfill this role, usually for issues that require a multi-disciplinary expertise (for example with regard to business and human rights, or discrimination against women) or that are more politically/culturally sensitive.

3. Finally, Special Procedures are well placed to undertake capacity-building activities to help developing countries apply human rights norms through legislation, policies, and practice.

Since the HRC's creation in 2006, many of the normative developments pertaining to new and emerging human rights challenges, including climate change and biodiversity, migration, poverty reduction, water and sanitation, and the human rights responsibilities of corporations, have been elaborated and promulgated through Special Procedures. If States are serious about ensuring that the technology sector operates consistently with international human rights objectives, they should again consider this proven institutional vehicle for building the normative edifice to guide that process. The mandate of such a Special Procedure could focus on ensuring that new tech products, including AI, serve (rather than undermine) human rights, and that all States—those in the Global North as well as those in the Global South—regulate such technologies in a manner consistent with their human rights obligations.

The principal counterargument to this strategy has to do with the efficiency of the Human Rights Council. As of November 2023 there are 60 Special Procedures mandates (46 thematic and 14 country mandates). The Universal Rights Group has consistently argued that this number should be reduced. The Council cannot meaningfully engage with the work of so many mandate holders. Moreover, those States that actively cooperate with the various Special Procedures are becoming overburdened by requests for country visits. Many observers also worry that the calibre of the various mandate-holders is decreasing. It is therefore understandable that some States are wary of calling for additional mandates to be created. The solution to the problem of too many mandates should not, however, be to institute a de-facto moratorium on the creation of new Special Procedures. The solution should rather be to rationalize existing thematic mandates, many of which no longer serve a useful purpose.

The Human Rights Council plays an important convening role in the promotion and implementation of global norms by States and technology companies. The HRC can convene intersessional roundtables or platforms for cooperation and dialogue (see below), designed to bring together States, UN experts, civil society, and technology companies to consider common solutions to the various human rights challenges and opportunities posed by digital technologies, including AI. The principal criticism of this idea is that it might potentially duplicate the existing groundbreaking work already underway in other parts of the UN system (e.g., the AI Advisory Body or the B-Tech Project). Such concerns miss two important points. First, the HRC has the clear mandate to serve as the primary and permanent forum in which human rights issues should be deliberated within the United Nations. While the HRC's remit extends only to human rights, it would be inappropriate to suggest that the forum should be closed to intersectional issues where human rights interrelate with other substantive domains (for example human rights and climate change, human rights and development, human rights and national security, or – in the instant case – human rights and new and emerging technologies).

Furthermore, as one of only three Councils at the UN, the HRC serves an important facilitator role. By convening a range of panels, roundtables and seminars, the HRC can bring together various institutional players across the UN system working on a given issue. In this way, the HRC can break through the disciplinary and institutional silos of the UN system and ensure greater policy coherence. Moreover, it can do so while also anchoring the overarching process in the broader language and logic of human rights. The HRC acting alone will never address the challenges and opportunities presented by the emerging digital future. It can, however, serve as a convener and catalyst for other stakeholders, working in concert, to do so using human rights as the overarching normative edifice to guide those efforts.

The Clarification, Distillation, and Presentation of novel human rights norms

The human rights system has only begun to grapple with the question of what norms should apply to the development and deployment of new and emerging digital technologies, and more fundamentally whether any new norms are needed at all. Initial responses by the human rights system have emphasized the adequacy of existing human rights norms, and the applicability in particular of the UNGPs to private efforts to develop trustworthy AI.

Various Special Procedures have also begun to grapple with the implications of AI on their mandate's particular focus. These efforts remain piecemeal, however, and depend on the initiative and capacity of existing mandate holders to engage with the impact of AI. As a result, States still lack a single integrated source for guidance on how they should craft human-rights based policies to govern new and emerging digital technologies. The work of OHCHR's B-Tech Project has moved in that direction, but still focuses primarily on big tech companies primarily located in the Global North, despite some recent efforts to also engage with smaller tech companies, including in the Global South. Moreover, the commendable work of the B-Tech Project does not diminish the value and need for intergovernmental consensus on how human rights norms should govern the development, deployment and regulation of AI.

If it wishes to continue engaging with various stakeholders active in the development of AI systems, the human rights system still needs to develop one single coherent, balanced, comprehensive, and overarching normative framework to address the human rights implications of AI. Such a framework should ideally focus on all human rights, discuss the needs of all vulnerable populations, and embrace both the need to guard against the risks of AI but also the need to harness AI as a driver for the progressive realization of human rights.

One option for the development of such a comprehensive normative framework could be the establishment of a requirement for all Special Procedure mandate holders to provide a periodic analysis of how AI-systems potentially affect their mandate. Another option would be to mandate an independent expert to compile framework principles on human rights and AI, based on similar work done by the Special Rapporteur on human rights and the environment or the former Special Rapporteur on Business and Human Rights. Such an approach would distill the various emerging normative initiatives into one coherent framework, speaking to a range of stakeholders. Such a mandate holder could also speak to the question of whether 'new' rights are needed to govern the promotion of human rights in the digital era, and build consensus on a core set of human rights principles to be applied in the pursuit of trustworthy AI.

The true added-value of a new special mechanism would be guidance on the concrete operationalization of aspirational human rights objectives throughout the AI product lifecycle (or more broadly speaking the 'datafication cycle' common to many new and emerging digital technologies). While the UN Working Group on Business and Human Rights and the B-Tech project have already made valuable contributions in this regard, a proposal for a single unified framework that could be endorsed inter-governmentally and by companies, for example as an annex to the UNGPs, would go a long way towards providing such guidance. In the view of the authors, such a mechanism would have to be a well-resourced, multi-stakeholder body with technical expertise in AI, business management and human rights law. Furthermore, such a mechanism would have to draw on expertise from a variety of different jurisdictions and disciplinary backgrounds. Such a mechanism could also be mandated to perform an appropriate monitoring functions by empowering it to communicate with stakeholders when their AI-related policies or practices jeopardize human rights.

Harnessing the convening power of the Human Rights Council

As described above, the HRC has already been quite aware of AI and its potential human rights implications. The HRC should build on this track record by convening multi-stakeholder forums in which to share best practices about the development and deployment of trustworthy AI. The HRC has at times struggled to get businesses involved in its activities. By embracing both efforts to mitigate the potential downside risks of AI as well as efforts to harness the upside potential of AI, the HRC can open a forum in which innovative social tech entrepreneurs are incentivized to showcase their ideas and participate in the work of the Council.

At a minimum, the Council could organize high-level multi-stakeholder panels or expert seminars to discuss best practices in human rights-based approaches to trustworthy AI. A precedent for such a workshop already exists in the form of a seminar organized by

the OHCHR in May 2020 on privacy and AI pursuant to HRC resolution 42/15.[349] Such multi-stakeholder discussion fora have the added benefit of advancing the assistive function of the Council, which is lacking in the context of the corporate responsibility to respect human rights, while promoting responsible business conduct. For example, the HRC could seek to take advantage of the clear brand competition between different approaches to trustworthy AI, such as the well-known competition between Anthropic and OpenAI, to help identify and clarify best practices in the field and help weigh in on those approaches from a human rights perspective. Rather than taking an overly broad approach, panels could consider specific human rights issues (e.g., AI and children's rights) or specific technical issues with human rights implications (e.g., algorithmic transparency) to identify best (or better) practices. Those better practices can then be disseminated and promoted, in the name of the Human Rights Council, as standards for other companies to emulate and exceed in their own product development efforts. Given the high visibility and perceived legitimacy of discussions hosted by the United Nations, such multi-stakeholder approaches would have the added benefit of furthering the Council's educational function by informing stakeholders of their rights and obligations as they relate to AI.

The Council should also continue to build on its essential mainstreaming function by promoting a human rights-based approach to new and emerging technologies across the entire UN system. The Council is well-positioned to convene various UN entities to discuss how AI-systems can affect their areas of work, and how their responses must be anchored in human rights. The Council should ensure that its work on trustworthy AI is disseminated throughout the UN system, most notably with the newly formed AI Advisory Board. Conversely, the Council, as a permanent body, should also request updates and briefings from other such bodies and processes.

Moving towards more institutional innovations, the Council might consider establishing a permanent platform where AI-tech start-ups claiming to have developed an AI-application with the potential to "make the world a better place" can showcase and promote their technologies. Such a platform would function in a similar manner as ITU's annual AI for Good conference. Companies would have the opportunity to demonstrate, using the logic and language of a human rights based approach (as opposed to corporate ethics) how their technologies or products can actively advance human rights. If convincingly made, the Human Rights Council can serve a function vaguely reminiscent of an "accelerator," launching innovative AI products designed to advance human rights into the marketplace with an added tailwind.

Finally, the Human Rights Council could potentially serve as a forum in which to gradually elaborate a global consensus about what is meant by "high risk contexts" in which AI should not be used. There have been various calls by technologists and human rights experts to ban the use of AI-systems in so-called "high-risk" contexts or when associated with supposedly "high-risk" technological capabilities. Examples of such "high-risk" use cases of AI that have been mentioned in the past include the use of AI in autonomous lethal weapons systems, in facial recognition technology, or in predictive policing. Few if any of these understandings of what constitutes a "high risk" use case for AI—and specifically which defining characteristics would distinguish a suspect or prohibited use case from a more legitimate and permissible one—enjoy any real consensus in the AI community. This is even more true when considering the divergent views of the industry, civil society, and government officials on these same questions. The Human Rights Council could serve the role of a consensus builder, working over time to distill our collective understanding of those use cases for AI technologies that are inherently problematic from a human rights perspective, and subsequently working towards appropriate and narrowly targeted collective international responses, such as moratoria or outright bans of certain AI use cases, which would be necessary to effectively tackle the risks given the transnational nature of such technologies. The Human Rights Council, drawing on its convening and normative functions, could host such discussions using a variety of tools at its disposal.

Driving implementation of a human rights-based approach and strengthening corporate accountability

Building on the need to strengthen corporate accountability, the HRC could also consider various avenues to raise the costs of irresponsible business conduct. For example, the Council could encourage Special Procedures to 'name and shame' egregious or irresponsible corporate behavior in their reports and interactive dialogues. Nothing precludes Special Procedures from requesting field visits to corporations in order to assess their policies and practice against international human rights law. While private companies may not wish to permit such a request, perhaps a combination of cross-industry visits to compare models and various non-disclosure stipulations could incentivize access. A dedicated Special Procedure could assess corporate policies and internal practices and publicly report on their alignment with international human rights standards. Such a Special Procedure would also be well placed to monitor regulatory developments and share best practices, with a view to ensuring State and corporate compliance with human rights obligations.

An even more ambitious proposal, which was first raised at the 9th Glion Human Rights Dialogue in May 2023, would be to create a mechanism mandated to assess at a technical level the compliance of AI systems with human rights standards. The proposal was to create a UN red team made up of software engineers and human rights experts to test systems to see if they have the potential to result in human rights harm, with a view to addressing such concerns at an early stage. Such a mechanism could also perform algorithmic audits to ensure that the contexts and purpose for which an AI-system is deployed align with human rights objectives. These are existing best practices that corporations already use to promote 'responsible AI', address bias and ensure responsible data collection and use. The difference would be that instead of management consultants, the UN could leverage its human rights expertise and moral authority and standing to raise the bar and hold businesses to a higher standard. Such a quasi-certification process could be made available either at cost or at a subsidized price to certain AI developers in the global south who might otherwise not be able to put in place similar safeguards.

Reinforcing State and non-State actor's capacity to align AI responses with human rights

There remains a clear and pressing need for the human rights system as a whole to improve its ability to assist other stakeholders solve problems. The UN system must increase its capacity and expertise to deal with complicated and technical AI-related issues and increase its capacity to provide concrete guidance on corporate policies and processes, as well as regulatory responses as they relate to the development and deployment of trustworthy AI. Enhancing the Human Rights Council's assistive function would correlate directly with the Secretary General's call in the proposed Global Digital Compact for member States to "establish[] a digital human rights advisory mechanism, facilitated by the Office of the United Nations High Commissioner for Human Rights, that would provide practical guidance on human rights and technology issues, building on the work of the human rights mechanisms and experts, showcase good practices and convene stakeholders to explore effective and coherent responses to legislative or regulatory issues."[350]

The most recent HRC resolution 53/29 on New and Emerging Technologies did allocate additional funding to OHCHR to increase its capacity in the field of AI. Mandating the establishment of regional technical advisors could help the Office monitor local developments and provide contextually relevant expertise to States to ensure that regulatory and legislative responses to AI are properly anchored in human rights.

These are moves in the right direction. The Human Rights Council and its associated Special Procedures should continue to play a role in the development of this budding institutional capacity, serving as a central hub for knowledge sharing and dissemination, as well as a source of guidance for those field officers providing technical support to Member States and interested corporate partners.

349. Human Rights Council, "The right to privacy in the digital age," (Oct. 7, 2019), UN Doc. No. A/HRC/RES/42/15, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/297/52/PDF/G1929752.pdf.

350. United Nations Executive Office of the Secretary-General (EOSG), "A Global Digital Compact — an Open, Free and Secure Digital Future for All: Our Common Agenda Policy Brief 5," (May 24, 2023), https://www.un.org/techenvoy/global-digital-compact.

Conclusion

It is said that humans are bad at recognizing and acknowledging exponential growth. The argument goes that we humans tend to imagine growth as a linear process. If I was able to walk 10 kilometers yesterday, I should equally be able to walk for 10 more today.

Growing numbers of AI researchers are telling us that this assumption of linearity is simply wrong, perhaps dangerously wrong, when it comes to AI. Their warnings are driving increasingly serious reflection and research in the technology sector on how to create trustworthy AI: an AI that does not produce unintended or unseen harmful side-effects, but also an AI that does not escape from human control.

These papers discussed three aspects of AI from a human rights perspective. The first looked at the very real efforts being made by entrepreneurs and technologists seeking to guard against the risks inherent to AI technologies. It focused its discussion on startups doing that work, looking at what is feasible for private corporations amidst their existential struggle for survival in an increasingly crowded market. What can startup entrepreneurs do, with the resources they have available, to nudge the products they create bend towards human rights?

Our findings suggest that there is quite a lot that startups can do when they embrace a human rights-based approach to technology. Small early investments tend to pay off handsomely. Short brainstorming sessions early in a startups' trajectory can yield major benefits in terms of corporate culture, internal guardrails, and public reputation that would be extremely difficult to cultivate in a much bigger corporation. As firms grow, and as they become more confident in their business model and the science allowing their products to reach their intended audiences, they can (and should) invest some more effort into ensuring that the internal processes exist that will continue to guide the startup in the right direction. Are robust and accessible grievance processes in place that will allow the company to understand when its products are having unintended or potentially harmful impacts? Has the company thought about how its products might impact vulnerable or marginalized communities, and has it invested efforts to mitigate those harms? These kinds of questions increasingly are standard business practice, and not only for supposedly 'progressive' firms. They are particularly important for startups working with AI. AI – for all the hype – still scares a lot of people concerned about their privacy, their livelihoods, and their personal sense of agency. Any tech company – big or small – that fails to soothe its customers' anxieties that their AI products may somehow turn into rogue robot killing machines will simply fail to succeed in the market. For the time being, therefore, tech companies large and small simply cannot afford to ignore the safety and trustworthiness of their products.

We also looked at the flipside of the Tech Paradox. Focusing on the issue of climate change, we looked at efforts by technologists and entrepreneurs to find creative ways to deploy AI to help solve climate change. Climate change serves as an interesting example to explore, both because it is one of the more pressing issues (including one of the more pressing human rights issues) of our time, but also because it offers us an illustration of the "upside potential" of AI to help solve some of humanity's most 'wicked' problems. These are problems that are so complex, and so interconnected with other issues and preconceived societal factors, that our usual scientific approach towards solving them simply fails. AI has the potential to vastly improve our collective ability as humans to understand and hopefully solve such problems.

Here too, we found that there is much that entrepreneurs can do to make the world a better place. We saw how corporations are investing efforts – often with no immediate commercial value to the company at all – that are making substantial contributions towards the fight against climate change. We also explored how these collaborations often hinge on new collaborations forming between businesses, governments, civil society organizations, and international institutions. These coalitions tend to be built on trust, dialogue, and exchange, relying as much or more on the "carrot" than the "stick" to encourage collaboration.

The third paper explored the rapidly evolving policy landscape at the United Nations, and in particular at the Human Rights Council. That discussion is itself also seemingly evolving at an exponential pace. What in late 2022 still seemed like an emerging topic of discussion, only one year later, in 2023, seems like an increasingly crowded space. AI and its impacts are being discussed today not only at the UN Human Rights Council, but also at the UN General Secretariat, the UNDP, UNESCO, OHCHR, the ITU, and countless other addresses. Our analysis shows the urgent need for a human rights frame to remain prominently represented in all of these conversations, underscoring that human beings enjoy a right to live a life of dignity, not a temporary privilege dependent on someone else's sense of personal or corporate ethics, charity, or good will. A few weeks before the release of our report, the UNOHCHR issued guidelines that strongly underscore and amplifies that point.

The question remains, now, where to devote our focus moving forward. The pace of AI's development will only continue to accelerate, as will our whiplashed discovery of both the potential risk and potential benefits of these new technologies. Much work remains to be done to find new, nimble, and effective ways for the UN, and the UN Human Rights Council in particular, to bridge the gap between their unrivaled institutional capacity to understand and highlight human rights, and the collective inability of non-technical experts to always understand the true impact of new and emerging technologies.

There is also an urgent need to think like a business consultant, however. In 2022, the first installation of this paper series proposed a set of 24 discrete processes that (we claimed) can be used by a variety of stakeholders – working in concert with each other – to drive forward a Human Rights-Based Approach to New and Emerging Technologies. In this year's paper series, we showed how those processes function in the AI sector, and specifically for startups. We discussed how a focus on broad and abstract principles – principles such as "transparency," or "legality," or "non-discrimination," while useful and philosophically fascinating, are perhaps less valuable from a practitioner's perspective as we might hope. This has also been the resounding conclusion of OHCHR's B-Tech Project report, which stands to become a cornerstone in the normative discussion about a human rights-based approach to new and emerging technologies.

Following that recommendation, therefore, we suggest that the next frontier for human rights practitioners is to develop concrete guidance, for the benefit of small entrepreneurs, or businesses operating in the Global South, to harness the best and brightest of what a management consultant might also have to offer to a paying client. How would a company think about institutionalizing a culture of constructive problem solving when confronted with a problem, for example? How should a technology company think about proactive transparency and disclosure? What role is there for a government seeking to regulate but not stifle technological innovation? How can civil society play a more active role in these discussions? These are the kinds of questions that we believe should be answered – in close consultation with a range of different stakeholders – and made available for anyone wishing to embrace a human rights-based approach to new and emerging technologies and its promise for humanity: harnessing new and emerging technologies safely, in a way that promotes greater human dignity without discrimination, empowers us to pursue our personal capabilities, and gives us the freedom to explore and discover new scientific and professional opportunities.

PERMANENT MISSION OF THE
REPUBLIC OF KOREA IN GENEVA

SNU AI POLICY INITIATIVE

UNIVERSAL RIGHTS GROUP

Not for sale
93360

9 791198 570314
ISBN 979-11-985703-1-4