Issue Brief

Decoding Intentions

Artificial Intelligence and Costly Signals

Authors
Andrew Imbrie
Owen J. Daniels
Helen Toner

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

October 2023

Executive Summary

How can policymakers credibly reveal and assess intentions in the field of artificial intelligence? AI technologies are evolving rapidly and enable a wide range of civilian and military applications. Private sector companies lead much of the innovation in AI, but their motivations and incentives may diverge from those of the state in which they are headquartered. As governments and companies compete to deploy evermore capable systems, the risks of miscalculation and inadvertent escalation will grow. Understanding the full complement of policy tools to prevent misperceptions and communicate clearly is essential for the safe and responsible development of these systems at a time of intensifying geopolitical competition.

In this brief, we explore a crucial policy lever that has not received much attention in the public debate: costly signals. Costly signals are statements or actions for which the sender will pay a price —political, reputational, or monetary—if they back down or fail to make good on their initial promise or threat. Drawing on a review of the scholarly literature, we highlight four costly signaling mechanisms and apply them to the field of AI (summarized in Table 1):

- *Tying hands* involves the strategic deployment of public commitments before a foreign or domestic audience, such as unilateral AI policy statements, votes in multilateral bodies, or public commitments to test and evaluate AI models;
- *Sunk costs* rely on commitments whose costs are priced in from the start, such as licensing and registration requirements for AI algorithms or large-scale investments in test and evaluation infrastructure, including testbeds and other facilities;
- *Installment costs* are commitments where the sender will pay a price in the future instead of the present, such as sustained verification techniques for AI systems and accounting tools for the use of AI chips in data centers;
- *Reducible costs* are paid up front but can be offset over time depending on the actions of the signaler, such as investments in more interpretable AI models, commitments to participate in the development of AI investment standards, and alternate design principles for AI-enabled systems.[1]

We explore costly signaling mechanisms for AI in three case studies. The first case study considers signaling around military AI and autonomy. The second case study examines governmental signaling around democratic AI, which embeds commitments to human rights, civil liberties, data protection, and privacy in the design, development, and deployment of AI technologies. The third case study analyzes private sector signaling around the development and release of large language models (LLMs).

Costly signals are valuable for promoting international stability, but it is important to understand their strengths and limitations. Following the Cuban Missile Crisis, the United States benefited from establishing a direct hotline with Moscow through which it could send messages.[2] In today's competitive and multifaceted information environment, there are even more actors with influence on the signaling landscape and opportunities for misperception abound. Signals can be inadvertently costly. U.S. government signaling on democratic AI sends a powerful message about its commitment to certain values, but it runs the risk of a breach with partners who may not share these principles and could expose the United States to charges of hypocrisy. Not all signals are intentional, and commercial actors may conceptualize the costs differently from governments or industry players in other sectors and countries. While these complexities are not insurmountable, they pose challenges for signaling in an economic context where private sector firms drive innovation and may have interests at odds with the countries in which they are based.

Given the risks of misperception and inadvertent escalation, leaders in the public and private sectors must take care to embed signals in coherent strategies. Costly signals come with trade-offs that need to be managed, including tensions between transparency for signaling purposes and norms around privacy and security. The opportunities for signaling credibly expand when policymakers and technology leaders consider not only whether to "conceal or reveal" a capability, but also *how* they reveal and the specific channels through which they convey messages of intent.[3] Multivalent signaling, or the practice of sending more than one signal, can have complementary or contradictory effects. Compatible messaging from public and private sector leaders can enhance the credibility of commitments in AI, but officials may also misinterpret signals if they lack appropriate context for assessing capabilities across different technology areas. Policymakers should consider incorporating costly signals into tabletop exercises and focused dialogues with allies and competitor nations to clarify assumptions, mitigate the risks of escalation, and develop shared understandings around communication in times of crisis. Signals can be noisy, occasionally confusing some audiences, but they are still necessary.

Table 1: Examples of Costly AI Signals

	Military AI and Autonomy	Democratic AI	Private Sector Signaling
Tying hands	Issue unilateral policy statements to convey intent, such as committing to maintain a human in the loop for nuclear command and control decisions.	Defend democratic AI principles by committing to predefined actions in response to AI-enabled adversarial attacks on democratic societies.	Release key information about advanced AI models, including transparency around the training data, model performance, and dangerous capabilities.
Sunk costs	Invest in red teaming procedures during training and before deployment and explore the use of emblems to facilitate attribution of AI-enabled weapons systems.	Release due diligence guidance for private companies operating in markets where there is a systemic risk of misuse of AI technologies.	Invest in trusted hosting services and test and evaluation infrastructure, including test beds and other facilities.
Installment costs	Commit to sustained verification techniques for AI-enabled systems and develop arrangements for intensive compute accounting.	Develop common certification standards, tools, and practices for AI auditors.	Commit to real-time incident monitoring and common standards around data collection and analysis of incidents involving AI-enabled systems.
Reducible costs	Set requirements and create incentives for investing in interpretable AI models and alternate design principles.	Sponsor prize competitions for AI safety research and the development of privacy-enhancing technologies that promote democratic values.	Publish AI impact assessments and the results of internal audits of AI systems

Table of Contents

Introduction

As the Cuban Missile Crisis neared its terrifying apex on October 22, 1962, Soviet First Secretary Nikita Khrushchev expressed dismay that his intended signal of deterrence had gone so awry. "Our whole operation was to deter the USA so they don't attack Cuba," the Soviet leader remarked to his inner circle.[4] With U.S. missiles in Italy and Turkey, he reasoned, why should the Soviets be denied the opportunity to right the balance? Khrushchev's decision to place missiles in Cuba was calculated to achieve a geopolitical trifecta: dissuade the Americans from invading the island, reestablish credibility at home, and seize the initiative from an increasingly assertive China. Moscow's motives were not readily apparent to analysts in Washington. Shortly after the Soviet launchers and missile shipments arrived in Cuba, an American U-2 reconnaissance plane captured evidence of the sites and relayed them back to a startled White House. U.S. President John Kennedy exclaimed to his advisors, "Why did he put these [missiles] in there...What's the advantage of that?"[5]

Against this backdrop of competing concerns and conflicting messages, a series of mishaps heightened tensions further. President Kennedy and Secretary of Defense Robert McNamara took pains to avoid what one historian observed was "the danger of having the Kremlin regard unauthorized actions as intentional 'signals.'"[6] On October 26, however, the U.S. Air Force conducted an intercontinental ballistic missile test at Vandenberg Air Force Base in California.[7] Then, on the morning of October 27, Soviet surface-to-air missiles struck an American U-2 spy plane in eastern Cuba, killing its pilot, Major Rudolph Anderson. Later that day, another American U-2 on a mission to collect samples of nuclear tests over the North Pole drifted into Soviet airspace without authorization. The U-2 maneuvered out of Soviet gunsights and returned home, but the risks of misperception were not lost on Washington. As a senior official from the State Department cautioned, "The Soviets might well regard this U-2 flight as a last-minute intelligence reconnaissance in preparation for nuclear war."[8]

The Cuban Missile Crisis is a reminder of the difficulty of sending clear and credible signals of intent in times of crisis. Leaders may think they are delivering one message, but the execution of their orders or lower-level actions of which they are unaware may convey another. Mirror imaging and the tendency to view other nations as monoliths only compound the challenge. Decades later, the United States once again confronts a world saturated with major power tensions, strategic arms competition, and the rapid advance of new technologies. The imperative to avoid miscalculation and communicate credibly is no less urgent today than it was during those 13 harrowing days in 1962.

Indeed, the task of signaling clearly may be even harder in the present environment. Innovation is more globalized and dispersed.[9] National security considerations increasingly permeate corporate decision-making on investment and supply chains.[10] Commercial players exert

influence on governmental decision-making, but, at times, act on the global stage independently or even against the national interest of their home countries.[11] Trust among the major powers has frayed and military-to-military communication has deteriorated.[12] Compounding matters, emerging technologies, such as artificial intelligence (AI), have become new playing fields for geopolitical competition.[13]

Advances in AI and machine learning, in particular, have altered the signaling landscape. Nations are vying for leadership over general-purpose technologies whose military and civilian applications are not easily differentiated.[14] AI algorithms and software services are intangible, though they are often tightly coupled with hardware components.[15] Such algorithms can be unpredictable in their effects and diffuse unevenly across sectors and societies. Openness has long characterized the academic field of AI, but concerns over safety and rising geopolitical and market pressures are accelerating the trend toward more closed ecosystems for AI development.[16] As the rivalry between the United States and China gathers momentum, the risks of mixed messages will grow as leaders broadcast the strengths of their AI-enabled systems and conceal weaknesses and intended use cases for deployment. Entanglement between nuclear and non-nuclear capabilities could raise the stakes even higher, as governments integrate AI into military decision-making and planning.[17]

In this context, it is critical that leaders pursue technology and national security policy goals without fueling instability or courting inadvertent escalation. The way forward will require a healthy dose of diplomacy and wise investments across a portfolio of standards, tools, and assessment approaches that facilitate responsible development across the life cycle of AI technologies.[18] One tool that holds promise but has received little attention in the public debate is what researchers have termed "costly signals." The essence of a costly signal is that the sender will pay a price if they back down or fail to make good on a promise or threat.[19] Costly signals reveal information of a certain type: governments or companies that send a costly signal are disclosing information that a less capable or resolved actor would not otherwise send.[20] The costs may be financial or reputational, or they may involve a cost in the human lives that such actions or statements put at risk, such as the deployment of troops to defend security commitments to allies.[21] For a signal to be costly and not a form of "cheap talk," the receiver must be able to observe compliance and the sender must be willing to risk paying a price for noncompliance.

Policymakers should be humble about the ability to convey accurate signals with critical and emerging technologies. Yet while signals can be noisy, they are still necessary. The solution is not to discount this important policy tool, but rather to wield it more effectively. Policymakers must understand the value and limitations of costly signals in AI and explore their potential applications for quickly advancing technologies that require careful net assessments of the cost, benefits, and risks for international stability.

This policy brief has four parts. Part one defines costly signals and why they matter in foreign policy. Part two outlines costly signaling mechanisms and maps them onto the field of AI to produce a framework of costly signals. Part three examines costly signals in practice by considering three case studies: major power signaling on AI-enabled weapons, U.S. government signaling on technology and democracy, and private sector efforts to signal restraint and responsible development and deployment of large language models (LLMs).[22] Part four draws out the policy implications and explores how and why costly signals may operate differently and elicit different reactions today than during the Cold War.

Costly Signals and Why They Matter

Policymakers rely on diplomacy and intelligence to gauge not only the capabilities of friend and foe, but also to discern their intentions. Information is at a premium, and leaders cannot discount the possibility that counterparts will bluff, mislead, or double deal to gain advantages over the other side. Is there any way out of this dilemma?

Researchers divide over two basic questions: whether leaders can divine intentions with any degree of certainty, and if so, whether statements or actions—words or deeds—are more dispositive of intent. Signaling pessimists argue that international relations are too uncertain, and the temptations to deceive are too great, for any signal of intent to be taken at face value.[23] Policymakers may be able to persuade friendly nations of benign motives, but interests can change in the future and no nation can conduct its foreign policy on the basis of lasting amity. By investing weight in the ability to shape their adversaries' intentions, leaders risk pursuing cooperative strategies with competitors that seem appealing in the near term but may leave them vulnerable over the long term.[24] Far wiser, pessimists argue, to assume the worst about other states' intentions and prepare accordingly.[25]

Signaling optimists, on the other hand, believe that intentions are discernable under certain conditions. The late theorist Robert Jervis distinguished between "signals" and "indices."[26] As he defined them, signals are "statements or actions" that are intended "to influence the receiver's image of the sender."[27] They are discrete actions that are observable, controllable, and inherently manipulable. As a result, they are telling but less reliable than what Jervis calls "indices," which are "statements or actions that carry some inherent evidence that the image projected is correct because they are believed to be inextricably linked to the actor's capabilities or intentions."[28] Indices are not under the control of the sender. They are useful on their own terms but also as a diagnostic for the signals and associated images that senders aim to present.

The distinction between signals and indices reflects a broader division among signaling optimists. Some argue that statements can be dispositive if they are delivered in private or threaten a rupture in ties.[29] Others claim that a signal's credibility is more closely tied to observable behaviors or shifts in material capabilities.[30] Still others point to institutional arrangements, domestic regime types, personal diplomatic impressions, and psychological traits as indicative of intent.[31] Evidence suggests that tying hands is not necessarily conditional on regime type.[32] In a democracy, accountability may take the form of losing an election; in competitive or closed autocracies where the leader relies on a clientelist group to stay in power, accountability may take more extreme forms.[33] While signaling optimists differ over the relevant variables, they share a common assumption: although intentions may be inconsistent, they are not inscrutable. Statements and behaviors can diverge, but they can also be tracked

over time based on a portfolio of indicators.[34] By understanding the context, operational concepts, and foreign policy dispositions of different leaders, states may form reasonable expectations about intent that can guide policymaking and mitigate the risks of accidents or inadvertent escalation.[35]

As governments and companies integrate AI into high-stakes systems that operate in increasingly complex environments, policymakers will need to understand the full range of tools at their disposal to reassure allies, restrain potentially threatening capabilities, and reveal intentions credibly. Costly signals can be an effective tool to achieve these goals, but it is important to understand the value and limitations of signaling in the rapidly advancing field of AI.

Costly Signaling Mechanisms and AI

Research on costly signaling offers a framework for thinking about intentions in the context of AI and machine learning. Based on a review of the literature, this brief elaborates on four signaling mechanisms: tying hands, sunk costs, installment costs, and reducible costs.[36] In practice, these mechanisms are not mutually exclusive. They can be employed in tandem to enhance the credibility of commitments and, at times, the lines between them blur. Taken together, they provide several avenues through which public, private, and non-governmental actors can signal intent on AI.

Tying hands involves the strategic deployment of public commitments before a foreign or domestic audience. The idea behind tying hands is that relevant audiences will hold a leader accountable if they do not make good on promises or threats. Suppose a leader pledges during a campaign to provide humanitarian aid to a stricken nation or the CEO of a company commits publicly to register its algorithms or guarantee its customers' data privacy. In both cases, the leader has issued a public statement before an audience who can hold them accountable if they fail to live up to their commitments. The political leader may be punished at the polls or subjected to a congressional investigation; the CEO may face disciplinary actions from the board of directors or reputational costs to the company's brand that can result in lost market share. In each case, the costs imposed are *ex post*, meaning they occur after the leader sends the signal, and they are *receiver-independent*, meaning they rely solely on the person sending the signal to make good on the promise or threat.[37]

In the context of AI, there are many examples of political leaders and companies employing the tying hands mechanism. U.S. military leaders have developed responsible AI principles and committed publicly and unilaterally not to cede decision-making on nuclear command and control to AI systems.[38] More recently, the U.S. Department of State issued a "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy."[39] Many companies have issued public statements and articulated AI principles to guide their decision making, with varying levels of transparency and accountability.[40] The company OpenAI sparked a vigorous public debate in 2019 when it announced that it would stage the release of its LLM, GPT-2, to avoid unintentional harm from misuse.[41] Since then, companies have experimented with a range of public release policies for their AI models.[42]

Beyond these examples, tying hands in AI could involve any number of policies and actions. Countries and companies could articulate public commitments in multilateral and multistakeholder fora that expose them to reputational costs or sanctions for noncompliance. Developers could pledge to adopt watermarking techniques in their products, commit to public evaluations and audits of their systems, and invest in assuring their AI models by generating evidence that they are sufficiently safe for their intended uses.[43] Private sector companies could

signal a commitment to data privacy by investing in privacy-enhancing technologies as well as smaller models and approaches that do not rely on massive pools of data.[44] Similarly, militaries could commit to unique emblems that facilitate attribution of AI-enabled systems.[45] Nations concerned about the risks of employing autonomous functionalities in weapons systems could sign up to codes of conduct that prohibit adversarial attacks on AI and machine learning resources or prescribe responsible conduct in certain areas of operation.[46] Such agreements could include voluntary pledges to accept third-party monitoring, common standards for test and evaluation procedures, and mechanisms to share information and resolve disputes.

Sunk costs rely on commitments whose costs are priced in from the start—unlike the tying hands mechanism, which involves public commitments that are only costly in the event of noncompliance. Similar to tying hands, however, sunk costs do not rely on the actions of the person receiving the signal. Sunk costs communicate a credible, often long-term commitment to a particular policy direction, buy-in from powerful stakeholders, and a lower likelihood of unexpected, drastic change from the set course. For example, one way a nation can indicate its resolve to use force is to mobilize large numbers of troops. The mobilization need not imply a decision to use force, but it is a costly signal that involves significant resources and political attention that cannot be recovered, which an otherwise irresolute nation may not send. In the context of AI and machine learning, sunk costs could include commitments to chain of custody requirements for advanced AI chips, licensing and registration of algorithms, and system inspections for AI verification, such as setting up verification zones to ensure that a system does not include AI chips or that AI chips are not controlling sensitive functionalities.[47] Nations and companies could commit large-scale investments for test and evaluation, including test beds and other facilities. A version of clinical trials for AI models could prove to be an equally costly signal that a company or public-private partnership is committed to transparency and responsible development. Virtual boundaries, or geofencing, and other design features could raise the costs up front and limit the capabilities or zones of operation of AI-enabled systems.[48]

Installment costs are costly commitments that the sender will incur in the future instead of the present. In contrast to the costs from tying hands, which are only incurred if the sender reneges on their commitments, installment costs are not reliant on the actions of the sender. They are fixed costs that cannot be recouped over time. For this reason, however, they can help extend the durability—not just the credibility—of commitments. Consider the costly signal of military basing arrangements. As research on costly signaling points out, the decision to establish a military base overseas engages two costly signaling mechanisms: significant investments up front (sunk costs) and a commitment to operate and maintain the base in the future (installment costs).[49] The time horizons and costly signaling mechanisms are related, but the logics differ in ways that have implications for assessing the credibility of commitments.[50]

As applied in the context of AI, installment costs could involve pledges by governments and companies to conduct risk assessments of AI models and make the results of those assessments available to the public. Governments could require, and private sector actors could implement, sustained verification techniques for AI systems, such as anti-tamper techniques that protect the integrity of software.[51] Given the important role of computing power in driving AI progress, policymakers and researchers are exploring compute accounting tools that track clusters of AI chips or specific properties of training runs in data centers for large models, such as the model weights or floating point operations per second above a certain threshold.[52] Efforts to codify and enforce these limits would leverage two costly signaling mechanisms: a costly public commitment to abide by the terms of the treaty (tying hands) and a longer-term commitment to intrusive monitoring and verification (installment costs).

Governments and companies can work together to signal credibly through installment costs. For example, governments could partner with companies to develop standardized practices, tools, and certifications for AI auditors.[53] Companies could work with governments to develop audit trails benchmarked against AI principles. They could also agree to provide data access for auditing purposes, involve relevant stakeholders in the process, and disclose the findings of audits publicly.[54] Contractual requirements between developers and deployers could include such requirements as costly signals of future intent. The Partnership on AI has developed an incident monitoring database based on voluntary input.[55] Publicly committing to standards for reporting incidents involving the use of AI models leverages installment costs by pledging transparency up front and then backing up that pledge with regular monitoring and evaluation.[56] Such an approach could support a more robust horizon scanning capability within governments and targeted regulations over time, including AI liability laws.[57] It could also help avoid misperceptions among rival nations. For example, governments could explore best practices for AI auditors and common standards around data collection and analysis of incidents involving AI-enabled systems.

Reducible costs are a final type of costly signal. In contrast to installment costs, reducible costs are paid up front but can be offset over time depending on the actions of the signaler.[58] For example, arms control agreements that provide for notifications of the movement of weapons systems or the collection and transmission of data on relevant forces and activities are costly future signals that can pay dividends to both sides in terms of greater transparency and stability.[59] In the AI context, reducible costs may take the form of private sector investments in more interpretable AI models and incentives for information sharing, such as model cards and data sheets that provide transparency on the training data, model weights, and other specific features of AI models.[60] It is costly for many companies to commit to such approaches unilaterally, but as AI models diffuse across societies and economies, companies

may recoup these costs over time by earning a reputation as a trustworthy and responsible developer of AI systems. Similarly, companies could develop investment standards for AI products and services that are consistent with the AI Principles of the Organisation for Economic Co-operation and Development (OECD).[61] The costs would be paid up front in terms of human capital development, financial resources, and dedicated staff time, but the benefits could be offset in the form of advantageous positions in supply chains and the ability to set the rules in competitive, next-generation markets.

As with other costly signaling mechanisms, governments and companies can work together to send costly signals of intent. Governments could promote responsible development by sponsoring prize competitions for AI safety and security or encouraging bounty programs for mitigating bias in AI systems.[62] Public-private partnerships could coordinate priorities and leverage shared resources for multilateral research and development initiatives on accident risks involving AI-enabled systems, including efforts to develop criteria for what constitutes an AI-related "incident" and best practices for the post-mortem process. Such cooperation could take the form of a Multilateral Artificial Intelligence Research Institute or international collaboration that draws lessons from the International Atomic Energy Agency or CERN, an intergovernmental organization for scientific research in fundamental physics.[63] Financial commitments and active contributions to a global research enterprise for AI safety could signal commitment to responsible AI development.[64] The startup costs would be significant, but governments and companies can recoup those costs by investing in AI safety research and best practices, thereby reducing the risks of accidents and inadvertent escalation.

In applying these costly signaling mechanisms, it is important to distinguish between the specific properties of AI models and the policy choices guiding their development and deployment.[65] Consider the challenge of understanding how an AI model "reasons" to make a prediction (sometimes called the "interpretability" problem). AI models can have billions of parameters, or "weights," that are updated based on large amounts of data or simulated environments where the model can infer decision rules through trial and error. The task of understanding which features of the training data mattered for a specific prediction is challenging.[66] Interpretability remains an active area of research in the field, but it already raises vexing questions in foreign and defense policy. Suppose an adversary were to deploy an AI-enabled system in combat to conduct intelligence, surveillance, and reconnaissance in contested waters. If the system mistakenly identified a merchant vessel as a naval ship and recommended kinetic strikes, how should the targeted government respond? On the one hand, the decision to strike was based, at least in part, on a faulty AI model deployed beyond the context for which it was trained. On the other hand, the target may not be privy to that information and will likely draw conclusions about the rival's intent based on its decision to deploy the AI-enabled system in the absence of safeguards.

A further complication in the signaling landscape is that not all actions are calculated to reveal intent. Companies may develop and deploy AI models for commercial reasons irrespective of the signal those decisions send to other states. Similarly, governments may impose regulations or take steps to accelerate innovation in AI for reasons unrelated to costly signals, even though such actions will affect how other states interpret their motives.[67] What's more, governments and companies conceptualize costs differently: governments may focus more on questions of national security and broader economic competitiveness and resilience, whereas companies will likely define costs in terms of market share and reputational constraints. Commercial players will also define costs based on where they are headquartered and their positions in global value chains. In short, domestic pressure groups, commercial interests, and governments respond to different political, social, and economic imperatives and pursue objectives that can be mutually reinforcing or conflicting depending on the context. As the case studies in this paper highlight, decisions that appear monolithic often reflect varying motives and time horizons among disparate actors.

Costly Signals in Practice

Military AI and Autonomous Weapons

If one wanted proof that it is hard to distinguish signals from the noise, a good place to start would be the international debate over lethal autonomous weapons (LAWS). This case study reveals the complexity of signaling in new and evolving areas of policymaking that concern not only government officials but also the statements and actions of commercial entities. Given the challenges of conveying intent in low-trust environments, this case explores the role of tying hands, sunk costs, installment costs, and reducible costs as mechanisms for stabilizing relations among the major powers as they compete to develop and deploy military AI applications.

Since 2014, nations have gathered in Geneva to develop principles for the potential use of such weapons.[68] Policymakers have debated where and how international law applies and the critical role of human judgment in the decision to employ autonomous weapons systems.[69] Both the United States and China have taken part in this process, and both countries have agreed to the consensus documents of the Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE), the United Nations body established in 2016 to examine issues related to these technologies. In 2019, the High Contracting Parties to the Convention on Certain Conventional Weapons adopted 11 guiding principles, including accountability, human responsibility, and the application of international humanitarian law to the development and potential use of LAWS.[70] Behind these consensus documents, however, lies substantial disagreement over the definition of autonomous weapons and the level of human involvement necessary to ensure compliance with international law. Since 2019, nations have struggled to reconcile these differences and momentum has stalled.

The challenge of signaling clearly and credibly is evident in China's 2016 and 2018 position papers submitted under the auspices of the GGE. In its 2016 position paper, China expressed concern about the ability of LAWS to adhere to the principles of distinction and proportionality under international law, noting that "such a weapons system presents difficulty in terms of accountability for its use."[71] While acknowledging the role of a new weapons review process, China made clear that it "supports the development of a legally binding protocol on issues related to the use of LAWS, similar to the Protocol on Blinding Laser Weapons, to fill the legal gap in this regard."[72] Two years later, however, China evolved its position. It enumerated five "basic characteristics" of LAWS, including lethality, autonomy, "impossibility for termination," indiscriminate, and evolution or the ability to "learn autonomously."[73] It concluded that "national reviews on the research, development and use of new weapons have, to a certain extent, positive significance on preventing the misuse of relevant technologies and on reducing harm to civilians."[74]

To U.S. observers, the differences between China's 2016 and 2018 position papers were ambiguous at best.[75] The definition of LAWS as lethal, irremediable, and indiscriminate in their effects would place them well beyond the pale of international law, and no responsible commander would seek to employ a weapon with such characteristics. By defining LAWS in the extreme but sanctioning the research and development of novel weapons with autonomous functionalities, China appeared to be implementing a principle of "legal warfare" to box in its competitors while creating flexibility for its own strategic imperatives.[76] Why shift from a position of public support for a legally binding protocol to a more equivocal stance on research and development if China did not want to pursue such a capability?

Irrespective of China's intentions for LAWS, U.S. analysts and policymakers have drawn conclusions from China's public statements and actions.[77] As one former Department of Defense (DoD) official testified before the U.S.-China Economic and Security Review Commission, "available evidence suggests that China is pursuing development of AI-enabled lethal autonomous weapons."[78] To bolster this claim, the former official pointed to China's definition of AI as a strategic priority in its 2017 New Generation AI Development Plan, in its 14th Five-Year Plan for 2021-2026, and in its most recent defense white paper. He also cited the statements of a senior executive at China's third-largest defense company. This executive expressed confidence that nations would continue to integrate AI and autonomy on the battlefield: "In future battlegrounds, there will be no people fighting."[79] Consistent with such statements, the former DoD official highlighted China's export of military unmanned systems and armed drones with autonomous functionalities, including Chinese military drone manufacturer Ziyan's Blowfish A2 model. He pointed to the company's website as claiming that the Blowfish A2 model "autonomously performs more complex combat missions."[80] The former official recognized the safety issues with AI-enabled weapons, but attributed the refusal of China's People's Liberation Army (PLA) to engage in defense policy dialogue with the DoD as evidence of its intent to develop LAWS and not be constrained by international norms.

Given the concerns over reliability and the risks of escalation with increasingly autonomous weapons, it is all the more important that nations send credible signals on LAWS. Yet doing so is challenging for three reasons. First, the technology is brittle and untested in battle. Unlike the production and assembly of nuclear weapons technology, military AI and autonomy are a fast-developing but nascent field of endeavor where the commercial sector plays a leading role. While the United States military has devised AI principles and updated its policy on LAWS, many nations have yet to clarify their national doctrine and processes for weapons with increasingly autonomous functionalities.[81] Modern AI systems are prone to accidents, opaque in their functioning, and can fail in ways that are surprising and hard to remediate.[82] Many AI models require training data that are specific to the context in which they will be deployed.

Data about relevant war-fighting domains is often incomplete, unavailable, or limited for reasons of security or legal and bureaucratic process. Countries are not fully transparent about their spending on LAWS, which makes it difficult to assess national-level capabilities and how those capabilities would perform during combat. As with debates over the regulation of AI in a domestic context, governments will face tradeoffs between AI model access, on the one hand, and concerns over national security and sensitive datasets, on the other.

Second, test and evaluation procedures for LAWS are underdeveloped and challenging to implement. Militaries must develop policy frameworks, standards, and metrics that are tailored to mission objectives. They must devise test and evaluation plans for AI-enabled systems that can learn and adapt over time in complex, dynamic operating environments, such as low-earth orbit or sub-surface locales.[83] Success is not easily defined, and the tradeoffs between safety and performance are hard to manage. Militaries must also guard against adversarial attacks and attempts to reverse engineer sensitive systems. As a consequence, test and evaluation for military AI systems will require continuous feedback between designers, developers, integrators, testers, and users. Militaries may also need to consider periodic retesting of AI-enabled systems even after deployment. Such approaches should focus not only on testing underlying algorithms but also on integrating AI software and hardware in a "system of systems" approach and developing human-machine frameworks that take into account cognitive biases and austere operating environments.[84] The willingness of countries to subject their systems to rigorous test and evaluation is unclear. Nations are pursuing military AI and autonomy under conditions of escalating geopolitical competition. The pressures to deploy untested systems for military advantage are ever-present, but they will grow more intense as countries mask relevant weaknesses in their programs and stoke distrust about their ultimate intentions.

Decoding signals on military AI and autonomy faces a third challenge: the increasing salience of commercial industry to defense innovation. Multinational corporations developing cutting-edge AI technologies may be headquartered in a single nation, but they are part of a global AI research enterprise with globalized supply chains. While their decisions can reflect national priorities, corporations are first and foremost subject to the demands of shareholders, financial markets, trade flows, and international economic trends. Compounding matters, AI is a general-purpose technology with a wide range of civilian and military applications. Partnerships between commercial entities and the government to develop dual-use technologies may end up supporting military innovation. China's efforts to develop a "techno-security state" that fuses its defense industrial base with civilian enterprises is well-documented, but the success of this strategy is difficult to measure.[85] Nonetheless, the close coupling of its military and civilian defense economies will encourage decision-makers to treat the statements and actions of Chinese commercial enterprises as indicative of national intent.

Indeed, one rationale for the U.S. decision in October 2022 to impose country-wide semiconductor-related export controls on China is the concern that dual-use technology partnerships with Chinese firms and civilian actors will be diverted to the PLA.[86] Chinese officials may also draw their own conclusions about DoD's efforts to strengthen cooperation with Silicon Valley and the growing ties between U.S. commercial entities and the U.S. military establishment.[87]

The increasing role of commercial industry in national security may enhance the credibility of commitments when public and private sector actors are in alignment, but it could also invite misperceptions when companies exaggerate their capabilities or take actions independent from their governments. For example, in the weeks following Russia's February 24, 2022, invasion of Ukraine, reports surfaced that Russia had deployed an AI-enabled drone to the battlefield.[88] As analysts observed, however, the weapon in question did not necessarily incorporate AI.[89] The Russian drone manufacturer and its parent company issued press releases that created ambiguity about the weapon's capabilities. The drone manufacturer, a subsidiary of the Russian arms maker Kalashnikov, claimed that the weapon could obtain coordinates from "[the sensor] payload targeting image."[90] Kalashnikov issued a separate press release boasting of the drone's AI-enabled capabilities for industrial and agricultural use cases. Neither of these two statements implies that the drone in Ukraine was equipped with AI to select and engage targets independently of human operators, but it would not be a stretch for governments to assume otherwise. Similarly, Ukrainians have operated the United Kingdom's Brimstone missile. The developer of this missile advertised several modes of operation, including a "fire-and-forget" mode that "provides through-weather targeting, kill box-based discrimination and salvo launch."[91] As experts were quick to point out, while the weapon likely operates in a semi-autonomous mode today, it is a software update away from potentially crossing the blurry threshold into a fully autonomous weapon.[92]

How can policymakers signal credibly in such complex operating environments? When it comes to LAWS, there are several mechanisms that governments and companies could leverage to communicate intent. Tying hands mechanisms offer one starting point. Just as the former head of the U.S. Joint AI Center Lieutenant General Jack Shanahan stated publicly that the United States would not integrate AI into nuclear command and control, governments could make unilateral policy statements on LAWS or enshrine such positions in official doctrine and processes.[93] One recent example is the United States' February 16, 2023, "Political Declaration on the Responsible Military Use of Artificial Intelligence and Autonomy."[94] While talk is cheap and public commitments can be walked back, unilateral statements of policy leave countries open to charges of hypocrisy and may entail reputational costs in the form of disapproving votes in multilateral bodies or lost support from friendly partners and domestic audiences, including the prospect of congressional investigation or budgetary restrictions.

The same logic could apply to America's competitors. With reports that Russia aims to deploy an autonomous, nuclear-armed underwater drone by 2027, the United States could urge China to make a unilateral statement of policy that such a capability would be destabilizing.[95] This signal would be costly for China, given its "no limits" partnership with Russia.[96] While Chinese leaders may decline to make a public statement to this effect, their refusal would send an important signal about China's relationship with Russia and potentially their own intentions to develop similar weapons, which would allow U.S. policymakers to update their assessments. Similarly, the United States, China, Russia, and other relevant countries and stakeholders could agree publicly to convene a series of Track 1.5 or Track 2 dialogues on AI safety.[97] These dialogues would be difficult to convene amid the onslaught of Russia's war against Ukraine. At the appropriate time, however, such conversations could not only surface potential areas of agreement on AI safety, but also clarify relevant national doctrine or policy related to LAWS as well as enhance transparency around the development and employment of military AI applications. Given public reports that China's PLA refused to discuss AI risk-reduction measures during the Defense Policy Coordination Talks of 2021, China could send a costly signal by allowing the PLA to participate in such dialogues and include this topic on the agenda.[98] By showing a willingness to define AI safety in practical terms and develop a common set of standards and testing protocols, the major powers could send a costly signal that they seek to reduce the risks of instability and inadvertent escalation.

The United States, China, and Russia could also explore sunk costs mechanisms. Nations could invest more and commit to transparency measures in test and evaluation procedures and allow relevant personnel to conduct site visits to test ranges and other facilities. Sharing safety technology will not necessarily make a competitor's system more effective. Indeed, evidence suggests that there can be tradeoffs between performance and safety.[99] The risks of improving the predictability of a competitor's AI-enabled systems must also be weighed against the benefits of reducing inordinate dangers to all sides.[100] Suppose Chinese leaders were to integrate AI more fully into their early warning systems. One does not need to rehearse the terrifying near-misses from the Cold War to know that such systems can be prone to failure in novel environments.[101] In a crisis scenario with the United States, would Chinese leaders regard such failures as unintended mishaps or preludes to an intentional attack, such as a conventional or nuclear counterstrike?[102] Given the relatively underdeveloped law, doctrine, and policy on incidents related to AI-enabled systems, a crisis involving such platforms could easily escalate to conflict.

Policymakers should also consider signaling with installment costs, or future costs that cannot be offset over time. The U.S.-Soviet Incidents at Sea Agreement of 1972 helped maintain stability and provided a mechanism for sharing information and resolving disputes.[103] As researchers have suggested, the major powers could sign an "International Autonomous

Incidents Agreement," which would invoke tying hands and installment costs as signals of intent.[104] Leaders could commit publicly to information-sharing and transparency measures or submit to intrusive monitoring and verification of their AI-enabled systems in designated geographic zones. Hardware inspections could verify whether AI chips are present in systems or controlling weapons functionalities.[105] Governments that commit to such measures publicly would send a costly signal about their intentions to abide by international norms in the development and potential use of LAWS.

Finally, governments could partner with industry leaders and university-affiliated research centers to implement reducible costs for AI-enabled military systems. Governments could set requirements and create incentives for investing in more interpretable AI models and alternate design principles, such as small data approaches to AI.[106] Policymakers and legislators could engage in public processes to develop common standards for military AI and explore the feasibility of sharing testing protocols with allies and competitors to mitigate the risks of escalation. As governments signal around the use of AI, they must be mindful that the technical characteristics of AI models can also confound efforts to send clear messages of intent. For this reason, policymakers should explore financial and other resource commitments to a global AI research enterprise charged with monitoring and measuring AI capabilities, improving methods for enhancing the interpretability of AI models, and developing a more robust empirical base for understanding and evaluating the dynamics of signaling in human-machine teams.

Democratic AI and Inadvertent Signals

Policymakers must keep in mind that both the intent of the sender and the predispositions of the receiver matter when it comes to sending and interpreting signals. Another important consideration involves audiences whom signalers may not be targeting but who nonetheless absorb public statements and declarations. This case study explores the implications of signaling around democratic AI development, regulation, and use (referred to with the shorthand of "democratic AI") for relationships with non-democratic partners. While much of the section focuses on government signaling, it also briefly examines the private sector's role in sending costly signals around democratic AI. The primary costly signal mechanism in evidence is tying hands, although this case study also highlights the role that installment cost and reducible cost mechanisms can play as part of the democratic AI toolkit.

Democratic AI has become a widely discussed topic in multinational fora and national AI statements. A broad definition of democratic AI based on these statements refers to AI applications that incorporate safeguards for democratic processes and societies into their development and deployment, as well as future democracy-protecting regulations. Examples include ensuring that systems are not biased against certain classes of citizens, whether by

poor data or algorithmic design; that governments do not use facial recognition or other potentially privacy-eroding AI applications in ways that infringe on citizens' civil liberties; and that adversaries and bad actors cannot use generative models to disrupt information environments to undermine faith in elections or the rule of law. This framing contrasts with authoritarian uses of AI, such as China's deployment of facial recognition and other AI applications in Xinjiang against the province's Uyghur ethnic minority, or censorship technologies and exploitation of data analytics with AI.[107]

Multinational and national-level government statements generally support this understanding of democratic AI, though they differ in the level of detail and specificity they provide. For example, at their 2023 summit in Japan, the G7 nations stated their determination to "advance international discussions on inclusive [AI] governance and interoperability to achieve our common vision and goal of trustworthy AI, in line with our shared democratic values."[108] The European Union's (EU) draft Artificial Intelligence Act, with new amendments adopted in June 2023, aims to promote "the uptake of human-centric and trustworthy artificial intelligence and to ensure a high level of protection of health, safety, fundamental rights, democracy and rule of law."[109] Other notable multilateral groupings calling for democratic values in the development and governance of AI include the OECD, Council of Europe, Global Partnership on AI, the United Nations Educational, Scientific, and Cultural Organization (UNESCO), the Freedom Online Coalition, and the U.S.-EU Trade and Technology Council, among others (see Appendix A).[110]

Individual national documents echo and, in some cases, expand on multilateral statements (see Appendix B). Australia, Brazil, Canada, Italy, New Zealand, Spain, the United Kingdom, and the United States are among the countries that have developed national AI strategies, principles, or vision documents that incorporate explicit considerations of democracy, though not all national statements focus on democratic principles and AI to the same extent.[111] For some, these statements reinforce multilateral declarations they have co-signed. The U.S. Blueprint for an AI Bill of Rights, created and adopted by the Biden administration, lays out five principles—safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation, and human alternatives consideration, and fallback—intended to protect society and ensure that AI progress does "not come at the price of civil rights or democratic values."[112] Other states, including France, Germany, Japan, and South Korea, have co-signed multilateral statements about democratic AI but do not mention them in their recent national documents.[113]

At present, democratic AI signals appear primarily intended to tie hands, indicating public commitments and sending messages against which leaders might one day be held accountable. The numerous multilateral and country-level statements mentioned above demonstrate hand-tying before foreign and domestic audiences. States have also borne some initial sunk costs in trying to organize and adopt democratic AI, such as the two U.S.-proposed

and co-organized Summits for Democracy. The Biden administration used the summits to tie hands, acknowledging the need for democracies to "put forward a vision of what they *stand for*—an affirmative, persuasive, secure and privacy-preserving, values-driven, and rights-respecting view of how technology can enable individual dignity and economic prosperity, and also what they will *stand against*," namely digital authoritarians' abuses of AI and other technologies.[114] In devoting resources, personnel, and capabilities to host the virtual summits, the United States and the summits' co-hosts also absorbed sunk costs they cannot immediately recoup to indicate their commitment to multilateral diplomacy around democratic AI.

In addition, statements and gatherings about democratic AI could result in longer-term installment costs or reducible costs as governments devote funding to democratic AI projects and hold future AI-enabled systems to formalized "democratic" standards. Legislation designed to protect democracy and democratic values from AI could create installment costs for governments that must enforce compliance with liability laws among public and private sector developers. The United States and United Kingdom jointly hosted a prize challenge with $3.75 million in awards for transatlantic AI developers who create privacy-enhancing technologies that reinforce democratic values, an example of a reducible cost whose benefits governments might reap over time by adopting the contest winners' creations.[115]

All of these costly signals about democratic AI, though mainly those intended to tie hands, appear geared toward communicating intentions to four general audiences: like-minded partners, domestic publics, the private sector, and authoritarian competitors. The message from the sender side is that governments intend to develop, encourage others to develop, and use AI in alignment with democratic values. The nuances differ for each audience.

Like-minded U.S. partners are clear receivers of signals about democratic AI, particularly when they are co-signatories of multilateral statements. They could interpret such signals as a desire to collaborate in areas of shared interest; alternatively, failure by a signatory to uphold previously agreed principles could result in reputational damage and diplomatic pressure from democratic peers.[116] Not all democratic governments strike the same balance in negotiating the tradeoffs between transparency around AI models for evaluation purposes and the goals of security, privacy, and data protection. Such differences between the United States and its allies create the opportunity for costly signals through the tying hands mechanism. Domestic audiences, including the general public, civil society groups, and the media, might use public commitments around AI principles to hold leaders accountable in the future. Journalists and interest groups—including researchers or think tanks, trade groups, and non-governmental organizations—could draw the public's attention to past statements if governments use or permit the development of AI that contradicts democratic values and civil rights, creating domestic political costs for leaders.[117]

The private sector, especially the tech industry, is a third key audience for these signals since governments are overwhelmingly consumers of AI technology and innovation from the commercial sector. Multilateral statements have even targeted the private sector, such as the "Call to the Private Sector to Advance Democracy" issued at the Summit for Democracy. The document appealed for greater commercial involvement in countering the misuse of technology and highlighted examples of how authoritarians and other actors have used technologies ranging "from machine learning models to surveillance technologies" to "polarize and fragment democratic societies . . . and erode public trust in democratic institutions," in addition to other harmful misapplications.[118] Governments signaling the importance of democratic values for AI development may expect private sector partners to incorporate these considerations into their system designs and consider refraining from selling AI technology to countries with poor human and civil rights records. For their part, firms may speak out when they are asked to develop AI capabilities, particularly for government stakeholders, that stand in opposition to democratic AI principles.

It is worth noting that the private sector, in addition to being an audience for government signals about democratic AI, may also send its own signals to consumers and other stakeholders. While occasionally referencing democratic AI in the same way as governments in such fora as the Summit for Democracy, commercial entities may also broadcast different interpretations of democratic AI, intentionally or not. For example, researchers from Google DeepMind published an article in the journal *Nature Human Behavior* entitled, "Human-Centered Mechanism Design with Democratic AI."[119] This paper focused not on electoral systems or processes, but instead on using AI to design redistributive economic policies "democratically" to benefit the most people at differing wealth levels.[120] Lack of clarity or shared definitions among government and private sector stakeholders around democratic AI, coupled with the private sector's leading technology development role, could make signaling on the topic in general more opaque.

A final audience is competitor states and near peers who might use AI-enabled or automated capabilities to attack the foundations of democratic societies, particularly election processes, or those who use AI to undermine human rights in their own societies. Threats of foreign interference in democratic processes using technology became particularly salient following Russian interference in the 2016 U.S. presidential elections and attempted interference by rogue actors in the 2017 French presidential election.[121] Recent advances in generative AI capabilities, including LLMs, have fueled concerns about the potential for adversaries to create and spread mis- and disinformation at scale.[122] Democratic AI statements and actions may therefore signal to Russia, China, and other competitors that the use of AI to attack democratic societies could engender a response. Though not directly related to AI, in 2020 then-candidate Biden vowed to "treat foreign interference in our election as an adversarial act that significantly

affects the relationship between the United States and the interfering nation's government," detailing retaliatory steps he would task his administration to take against a foreign meddler.[123] U.S. Secretary of Defense Lloyd Austin stated that "our use of AI must reinforce our democratic values, protect our rights, ensure our safety, and defend our privacy" against the AI "pacing challenge" of China.[124]

Given the signals policymakers aim to send to these different audiences, the framing of democratic AI, particularly in opposition to authoritarianism, may be a useful shorthand for distinguishing the approaches of democratic nations from those of competitors. Yet this framing belies the more complicated reality that democratic states frequently collaborate with authoritarian governments to protect their own interests and security. Furthermore, democracies often defend such cooperation by underscoring the need to firm up relationships with global swing states amid competition with China.[125] The United States has a broad network of global partners ranging from weak democracies to undemocratic and authoritarian states, many of whom might be uninterested in or even opposed to technology developed according to democratic values. Statements about democratic AI alone may not necessarily push them closer to China, but where the quality of democratic- and authoritarian-developed AI is comparable, non-democratic partners may choose to adopt the latter set of technologies with no strings attached.[126] Democratic policymakers should not abstain from trumpeting democratic principles on these states' accounts, but they should consider the potential consequences of statements about democratic AI if they choose to rely on these partners in the future.

The monarchies of the Gulf Cooperation Council (GCC) offer examples of authoritarian U.S. partners for whom associating democracy with AI could create diplomatic and strategic challenges and negatively impact security. Saudi Arabia, the United Arab Emirates (UAE), Qatar, Bahrain, Kuwait, and Oman have individually and collectively cultivated strategic relationships with the United States, premised on a long-standing American security guarantee in exchange for cooperation on energy and security interests.[127] Today, the GCC states host more than 30,000 U.S. military personnel, multiple U.S. Central Command (CENTCOM) headquarters across military domains, multinational maritime task forces, and they provide access to at least 20 basing facilities throughout the Gulf.[128] Cooperation in the past two decades of U.S. operations in CENTCOM has featured intelligence sharing, assistance in political negotiations, and even some joint counterterrorism operations.

Despite their significance, U.S.-Gulf relations have been difficult and even fractious. Tensions stem from differing policy and threat assessments to legitimate U.S. concerns around the suppression of dissent, civil liberties, and women's, minorities', and migrant workers' rights in the Gulf, among others. U.S. lawmakers and civil society have led high-profile criticism and calls for the United States to distance itself from these partners, particularly Saudi Arabia.[129]

Furthermore, while their most significant security partner remains the United States, China is a leading Gulf trade partner, complicating U.S. efforts to rely on the GCC states amid technological and strategic competition.[130] U.S.-Gulf cooperation persists, but often in spite of a challenging misalignment of political systems, values, and, sometimes, interests.

The Gulf states are worth examining because of their role in intelligence, basing, and access partnerships and because their adoption of non-democratic AI systems, particularly those developed by China, could impact U.S. security. The long history of U.S.-Gulf relations may suggest that the GCC states do not see democratic messages as applicable to them. However, costly signals about democratic AI complicate this dynamic. Since the United States is signaling that democratic AI will impact the design and deployment of particular technologies, the reactions of Gulf partners to messaging about values may turn on how and whether they believe that technology with democratic values "baked-in" serves their interests. In this context, exploring how Gulf partners might react to inadvertent U.S. signals about democratic AI and the AI capabilities they might adopt is instructive, given the potential national security implications.

One possibility is that democratic AI signals could have little impact on Gulf partners or be dismissed by them as cheap talk. They could interpret democratic AI signals as extensions of U.S.-China competition, rather than indicative of a differentiated, values-based approach. Gulf partners could buy the best technology they are able to access, regardless of who develops it, leaving democratically developed AI to compete with authoritarian technology on cost and technical merits. In this case, democratic AI might not necessarily dissuade Gulf partners from purchasing U.S. technology, but could exacerbate strained political and diplomatic relations.[131] Another possibility is that Gulf partners might refrain from buying certain U.S. AI products and services they could use for surveillance applications, such as facial recognition and data analytics, if they interpret from U.S. signaling that such products and services are designed with democratic safeguards and unlikely to help them address regime security concerns.[132] Efforts to counter the proliferation of AI capabilities used for autocratic purposes would align with U.S. national and multinational democratic AI commitments. However, such commitments would provide the United States scant leverage to dissuade partners from buying these capabilities from China. This outcome could, in turn, deepen U.S. worries about China's growing regional influence and U.S. network and intelligence security.[133]

The experience of 5G adoption in the Middle East with Huawei offers insight into how authoritarian partners in the Gulf may respond when they do not perceive the United States as a reliable provider of a strong technological alternative. The United States previously expressed concerns to Saudi Arabia, the UAE, and Bahrain in 2019 over the installation of Huawei's 5G telecommunications infrastructure. Officials and elected representatives communicated the potential negative impact on intelligence sharing for countries adopting

Huawei's technology.[134] Nonetheless, the Gulf's largest telecom providers reached agreements to develop 5G networks in partnership with Huawei to fulfill national modernization plans, such as Saudi Vision 2030.[135] The Gulf states have since decreased their exposure to U.S.-China tensions around 5G by investing in Open RAN systems, allowing feasible alternative 5G providers to Huawei to enter their markets.[136] They have not, however, severed ties with Huawei to the same extent as Europe.[137] Reporting in 2023 cited the UAE's Huawei deal as one indicator of close ties to China holding up F-35 aircraft and MQ-9 drones sales from the United States.[138] If Gulf partners begin to incorporate Chinese-developed AI into their systems on the basis that they are uninterested in using democratic AI, it could heighten U.S. concerns about data security and interoperability. Such concerns may even lead to reduced intelligence-sharing. Gulf partners' adoption of Chinese technology could also further enhance China's ability to lead AI standards development in applications useful for authoritarian regimes, such as facial recognition.[139]

Outside of the Persian Gulf and beyond security issues, the United States has a number of strategic and economic non- or weak democratic partners who may bristle at democratic AI messaging. For example, Singapore is developing its own significant AI ecosystem by building domestic talent and attracting foreign investment from both the United States and China.[140] As the United States competes with China to access Singapore's AI market, democratic signaling could create uncertainty with the country's government that puts the United States at a comparative disadvantage relative to China. The implications for strategically important but democratically backsliding nations, such as India, will also need to be managed carefully.[141]

Finally, the United States may be exposed to charges of hypocrisy or moral compromise for dealing with authoritarian partners and undercutting its democratic values.[142] This challenge has long bedeviled U.S. ties with the Gulf countries and could do so with other undemocratic nations. Given the United States has stressed the importance of democratic AI development, however, creating technology partnerships with non-democracies or sharing capabilities could provoke a backlash from domestic stakeholders and other democratic partners. The view that the United States might be supporting authoritarian applications of AI abroad, even if only through allowing private companies to provide technology to non-democratic regimes, could undermine the credibility of U.S. and allied signaling about democratic AI's importance.

The United States and other like-minded nations should not refrain from laying out principles to guide the development of AI that align with closely held democratic values. The task of articulating a positive vision for democratic AI is important, as is the process of establishing rules of the road that protect the sanctity and legitimacy of democratic processes, including election integrity, protection against mis- and disinformation, and safeguards for civil liberties and human rights. The defense of these values is worth the diplomatic costs. Yet the United States has many non-democratic partners, and non-aligned and global swing states may be

unsure of how to interpret democratic AI signals that are not necessarily targeted at them.[143] Policymakers should consider the broad range of audiences who may be receiving the signals they broadcast and take into account how this diversity of perspectives may complicate the messages they are trying to convey at home and abroad.

Private Sector Signaling

A notable feature of the present era is that—unlike during much of the 20th century—strategic technologies are no longer primarily developed in laboratories run or funded by governments. AI is no exception, with many of the most advanced systems being developed in consumer-facing technology companies. This shift in the center of gravity of where technologies are developed means that governments and the private sector are deeply interwoven and relevant signals could be sent by an expanded set of actors. As the case studies on signaling around lethal autonomous weapons and democratic AI show, observers seeking to anticipate the trajectory of AI development and use must now attend not only to signals from governments, but also from a range of industry players who increasingly contribute essential functions and services in conflict environments, such as the ongoing contributions of major tech platforms in Ukraine.[144]

The growing role of private sector entities in national security underscores the complexity of the signaling landscape and the challenges involved in reducing misperceptions and miscalculations amid geopolitical tensions. To better understand the dynamics around signaling in a commercial context, the case studies laid out below provide two different examples of companies sending costly signals of their intentions to develop technology safely and responsibly. The first case examines the role of tying hands and reducible costs as signaling mechanisms. The second case explores how companies can leverage installment costs to convey intent and strengthen norms around the release of potentially destabilizing capabilities.

A long-standing concern among analysts of AI development is the possibility of a "race to the bottom," in which multiple players feel pressure to neglect safety and security challenges in order to remain competitive. Perceptions—and therefore signals—are key variables in this scenario. Most actors would presumably prefer to have time to ensure their AI systems are reliable, but the desire to be first, the pressure to go to market, and the idea that competitors might be cutting corners can all push developers to be less cautious.[145] Accordingly, signaling has an important role to play in mitigating race-to-the-bottom dynamics. Parties developing AI systems could emphasize their commitment to restraint, their focus on developing safe and trustworthy systems, or both. Ideally, credible signals on these points can reassure other parties that all sides are taking due care, mitigating pressure to race to the bottom.

Much private sector signaling on AI speaks directly to these concerns. The highest levels of leadership at major tech companies have emphasized the importance they place on building safe and trustworthy systems. Microsoft president Brad Smith described his firm as "committed and determined as a company to develop and deploy AI in a safe and responsible way," while Google CEO Sundar Pichai stated that "we are taking our time to [perform safety checks], and we'll continue to be very, very responsible."[146] As with the public commitments discussed earlier in this paper, these broad statements reflect one approach to costly signaling.

To more fully understand how private sector actors can send costly signals, it is worth considering two examples of leading AI companies going beyond public statements to signal their commitment to develop AI responsibly: OpenAI's publication of a "system card" alongside the launch of its GPT-4 model, and Anthropic's decision to delay the release of its chatbot, Claude. Both of these examples come from companies developing LLMs, the type of AI system that burst into the spotlight with OpenAI's release of ChatGPT in November 2022.[147] LLMs are distinctive in that, unlike most AI systems, they do not serve a single specific function. They are designed to predict the next word in a text, which has proven to be useful for tasks as varied as translation, programming, summarization, and writing poetry. This versatility makes them useful, but also makes it more challenging to understand and mitigate the risks posed by a given LLM, such as fabricating information, perpetuating bias, producing abusive content, or lowering the barriers to dangerous activities.

In March 2023, California-based OpenAI released the latest iteration in their series of LLMs. Named GPT-4 (with GPT standing for "generative pre-trained transformer," a phrase that describes how the LLM was built), the new model demonstrated impressive performance across a range of tasks, including setting new records on several benchmarks designed to test language understanding in LLMs. From a signaling perspective, however, the most interesting part of the GPT-4 release was not the technical report detailing its capabilities, but the 60-page so-called "system card" laying out safety challenges posed by the model and mitigation strategies that OpenAI had implemented prior to the release.[148]

The system card provides evidence of several kinds of costs that OpenAI was willing to bear in order to release GPT-4 safely. These include the time and financial cost of producing the system card as well as the possible reputational cost of disclosing that the company is aware of the many undesirable behaviors of its model. The document states that OpenAI spent six months on "safety research, risk assessment, and iteration" between the development of an initial version of GPT-4 and the eventual release. Researchers at the company used this time to carry out a wide range of tests and evaluations on the model, including engaging external experts to assess its capabilities in areas that pose safety risks. These external "red teamers" probed GPT-4's ability to assist users with undesirable activities, such as carrying out cyberattacks, producing chemical or biological weapons, or making plans to harm themselves

or others. They also investigated the extent to which the model could pose risks of its own accord, for instance through the ability to replicate and acquire resources autonomously. The system card documents a range of strategies OpenAI used to mitigate risks identified during this process, with before-and-after examples showing how these mitigations resulted in less risky behavior. It also describes several issues that they were not able to mitigate fully before GPT-4's release, such as vulnerability to adversarial examples.

Returning to our framework of costly signals, OpenAI's decision to create and publish the GPT-4 system card could be considered an example of tying hands as well as reducible costs. By publishing such a thorough, frank assessment of its model's shortcomings, OpenAI has to some extent tied its own hands—creating an expectation that the company will produce and publish similar risk assessments for major new releases in the future. OpenAI also paid a price in terms of foregone revenue from the period in which the company could have launched GPT-4 sooner. These costs are reducible in as much as OpenAI is able to end up with greater market share by credibly demonstrating its commitment to developing safe and trustworthy systems. As explored above, the types of costs in question for OpenAI as a commercial actor differ somewhat from those that might be paid by states or other actors.

While the system card itself has been well received among researchers interested in understanding GPT-4's risk profile, it appears to have been less successful as a broader signal of OpenAI's commitment to safety. The reason for this unintended outcome is that the company took other actions that overshadowed the import of the system card: most notably, the blockbuster release of ChatGPT four months earlier. Intended as a relatively inconspicuous "research preview," the original ChatGPT was built using a less advanced LLM called GPT-3.5, which was already in widespread use by other OpenAI customers. GPT-3.5's prior circulation is presumably why OpenAI did not feel the need to perform or publish such detailed safety testing in this instance. Nonetheless, one major effect of ChatGPT's release was to spark a sense of urgency inside major tech companies.[149] To avoid falling behind OpenAI amid the wave of customer enthusiasm about chatbots, competitors sought to accelerate or circumvent internal safety and ethics review processes, with Google creating a fast-track "green lane" to allow products to be released more quickly.[150] This result seems strikingly similar to the race-to-the-bottom dynamics that OpenAI and others have stated that they wish to avoid. OpenAI has also drawn criticism for many other safety and ethics issues related to the launches of ChatGPT and GPT-4, including regarding copyright issues, labor conditions for data annotators, and the susceptibility of their products to "jailbreaks" that allow users to bypass safety controls.[151] This muddled overall picture provides an example of how the messages sent by deliberate signals can be overshadowed by actions that were not designed to reveal intent.

A different approach to signaling in the private sector comes from Anthropic, one of OpenAI's primary competitors. Anthropic's desire to be perceived as a company that values safety shines

through across its communications, beginning from its tagline: “an AI safety and research company.”[152] A careful look at the company’s decision-making reveals that this commitment goes beyond words. A March 2023 strategy document published on Anthropic’s website revealed that the release of Anthropic’s chatbot Claude, a competitor to ChatGPT, had been deliberately delayed in order to avoid “advanc[ing] the rate of AI capabilities progress.”[153] The decision to begin sharing Claude with users in early 2023 was made “now that the gap between it and the public state of the art is smaller,” according to the document—a clear reference to the release of ChatGPT several weeks before Claude entered beta testing. In other words, Anthropic had deliberately decided not to productize its technology in order to avoid stoking the flames of AI hype. Once a similar product (ChatGPT) was released by another company, this reason not to release Claude was obviated, so Anthropic began offering beta access to test users before officially releasing Claude as a product in March.

Anthropic’s decision represents an alternate strategy for reducing “race-to-the-bottom” dynamics on AI safety. Where the GPT-4 system card acted as a costly signal of OpenAI’s emphasis on building safe systems, Anthropic’s decision to keep their product off the market was instead a costly signal of restraint. By delaying the release of Claude until another company put out a similarly capable product, Anthropic was showing its willingness to avoid exactly the kind of frantic corner-cutting that the release of ChatGPT appeared to spur. Anthropic achieved this goal by leveraging installment costs, or fixed costs that cannot be offset over time. In the framework of this study, Anthropic enhanced the credibility of its commitments to AI safety by holding its model back from early release and absorbing potential future revenue losses. The motivation in this case was not to recoup those losses by gaining a wider market share, but rather to promote industry norms and contribute to shared expectations around responsible AI development and deployment.

Yet where OpenAI’s attempt at signaling may have been drowned out by other, even more conspicuous actions taken by the company, Anthropic’s signal may have simply failed to cut through the noise. By burying the explanation of Claude’s delayed release in the middle of a long, detailed document posted to the company’s website, Anthropic appears to have ensured that this signal of its intentions around AI safety has gone largely unnoticed. Taken together, these two case studies therefore provide further evidence that signaling around AI may be even more complex than signaling in previous eras.

Policy Considerations and Lessons Learned

Costly signals offer a way to communicate intentions in situations of low trust, but they operate differently today than during the Cold War. The economic context has transformed, and the role of commercial entities in driving innovation has expanded significantly. Dual-use technologies present challenges and opportunities for messaging clearly in an increasingly contested global science and technology landscape. Based on a close examination of major power signaling on military AI and autonomous weapons, U.S. government signaling on democratic AI, and private sector signaling around the release of powerful language models, this study highlights the following policy considerations and lessons learned.

Signals are not as "loud and clear" as they once were. Policymakers during the Cold War experienced no shortage of nuclear crises fueled by misperceptions, but there are limits to comparing costly AI signals with diplomacy around nuclear weapons technologies. The scope and scale of AI's commercial impact is vastly larger and the resource base is both more concentrated (in the case of advanced chips and the photolithography equipment used to make them) and more diffuse (in the case of open-source data and AI software). The post-Cold War period has seen the rise of non-governmental actors, each with varying degrees of influence on models for AI governance and the contemporary signaling landscape. Policymakers must also contend with the growing national security implications of general-purpose technologies, such as AI and advanced node semiconductors. It is not easy to distinguish between the military and civilian uses of such technologies. Doing so requires expertise, significant resources, technical infrastructure, and global situational awareness of science and technology trends.

The economic entanglement of nations further complicates the signaling picture. Despite pressures toward supply-chain reshoring and "de-risking" of critical and emerging technologies in select areas, countries and companies remain deeply interconnected in today's global economy. Governments and private sector actors can leverage complex economic and financial networks and supply chains to send costly signals by restricting or expanding capital flows, approving or denying foreign investment, and imposing or lifting trade controls.[154] At the same time, the increasing role of private sector companies in driving innovation creates challenges for sending clear signals of intent in AI. Policymakers must interpret multiple, often conflicting, signals from governments and private sector actors that may not share the same information, conception of costs, or geographic location. Such "noisy" environments present obstacles for signaling, but they can also create opportunities.[155] By dispatching multiple signals and gauging the reactions of target audiences, leaders can adjust their messaging to amplify those signals that achieve the intended effect.

Signals can be inadvertent yet potent. The distinction between intentional and unintentional signals highlights the growing complexity of the signaling landscape for policymakers. Not all

signals fall within the purview of government officials, and actions intended to convey one message may resonate differently with foreign and domestic audiences. U.S. government messaging on technology and democracy is a form of inadvertent costly signaling. This posture risks straining ties with partners who may not share these values, such as countries in the Gulf Cooperation Council, the Group of 77 in the United Nations General Assembly, and partners in Southeast Asia. Many of these governments pursue hedging strategies between the United States and China to maximize their autonomy in an increasingly competitive international environment. U.S. government messaging on technology and democracy could encourage partners to tilt toward China's no-strings-attached commercial approach to technology development and away from the United States' commitment to values-based design. U.S. government signaling is costly in another way: it leaves the U.S. government open to charges of hypocrisy for articulating support for technology and democracy and then partnering with countries that do not share these values.

Costly signals are only one tool in the AI policy toolkit and must be embedded in comprehensive strategies. The leading role of commercial firms in AI development underscores the need for coordinated actions and strong partnerships between the public and private sectors. As the first case study highlighted, there is a distinction between the technical characteristics of AI models and the policies that shape their design, development, and use in a military context. Governments have more influence on the latter than the former, though both sides of the equation have implications for how rival states will interpret costly signals. Policymakers may decide to deploy AI-enabled systems that meet certain thresholds for safety.[156] While governments control the decision to deploy such systems, they can only indirectly influence the course of technical research and progress on robustness and interpretability in the field of AI. The signaling logics differ, but rival states may not distinguish between the concerted decisions of governments and faulty AI-enabled systems that are deployed beyond the context for which they were trained. The second case study examined the strengths and limitations of costly signaling in a competitive context where the sides may be pursuing different objectives. In such environments, messages are not always relayed or interpreted in the manner policymakers assume. If companies in the United States or allied countries design and sell AI-powered surveillance capabilities abroad, for example, such actions can undermine the signals policymakers think they are sending on technology and human rights.[157]

The policy choice is not simply whether to conceal or reveal AI capabilities, but also how to reveal them and through which channels. Signals in AI can be costly in different ways. Test and evaluation approaches for AI-enabled weapons will signal different messages depending on the degree of transparency and whether the focus is on civilian or military test and evaluation procedures, and whether they include sharing technologies and joint access to test

ranges and infrastructure. The content and channels of the message matter and will add a layer of complexity to the signals a party aims to convey. Concurrent signaling from public and private sector actors may indicate greater clarity of purpose and resolve than divergent or multivalent signaling. The political context also matters. Misperceptions about what counts as authoritative in the political context of a rival nation may confound signaling attempts or communicate intent in ways that have unintended consequences.

Signals are an indelible part of the contemporary foreign policy landscape, so it is worth examining how policymakers can communicate clearly and avoid misperceptions. One path forward is for governments to leverage procurement practices and regulations to shape norms around AI development and use.[158] For example, policymakers could work with industry experts and academic researchers to enshrine norms around AI transparency (such as the release of model cards, system cards, or similar documentation) through procurement policies, including appropriate protections for privacy and security. The complexities involved in signaling would also benefit from focused Track 1.5 dialogues and table-top exercises among U.S. allies and competitor nations. Scenario-based exercises would provide governmental and non-governmental actors the opportunity to stress-test assumptions and better understand how different parties conceptualize signals, define costs, and manage the risks of escalation. By incorporating signaling into policy dialogues between allies and competitors, policymakers could facilitate the development of norms and shared understandings around signaling in different contexts and at various levels of escalation.

The coupling of public and private sector messaging and actions can be a powerful source of multivalent signaling. Signals can come from multiple voices and sources. This form of multivalent signaling can enhance the credibility of commitments when the signals are aligned and come from two or more independent actors. Multivalent signaling can also complicate the task of messaging clearly. The first case study demonstrates the challenges of signaling on AI-enabled weapons, particularly when public and private sector actors send divergent signals or when policymakers interpret the signals of private sector actors as indicative of national intent. Companies in freer markets may respond to national priorities, but they are also more accountable to shareholders, financial markets, and global economic trends as compared with national champions in authoritarian states. Profit motives may encourage some businesses to exaggerate their capabilities or send signals at inopportune moments. Some governments may leverage the ambiguity of noisy signaling environments to claim plausible deniability for adverse outcomes generated by private sector actions or statements. In short, the time horizons of the battlefield and boardrooms are not always aligned.

As a tool of technology policy, costly signals come with their own trade-offs that need to be managed.[159] The cases in this paper highlight the tensions between transparency for signaling purposes and norms around privacy and security. External audits of AI algorithms

and greater transparency around the data used to train large models are features, not bugs, of a safe and responsible approach to AI development. External audits enable third parties to corroborate internal test and evaluation procedures and surface areas of public concern that are not within the immediate field of vision of private sector actors.[160] In practice, however, external audits may reveal personal data or expose proprietary information about algorithms that put companies at a disadvantage. More information about AI systems can also overwhelm consumers and widen the attack surface for unscrupulous actors who seek to exploit vulnerabilities of AI models or the larger systems of which they are a part. Researchers are exploring the use of query-based approaches and structured transparency as methods for resolving the tensions inherent in external audits of AI systems.[161] Technical approaches show promise for managing these trade-offs, but policymakers will also need to explore creative institutional, policy, legal, and regulatory mechanisms to balance concerns among parties across the life cycle of AI development.

The ability to convey costly, credible, and clear signals may vary depending on the context and technology area. Critical and emerging technologies have different characteristics and requirements that may expand or constrain the scope for costly signaling. For signaling purposes, it is helpful to think of critical and emerging technologies along a spectrum based on their capital expenditures, controllability, and covertness.[162] Capital expenditures impact the number of actors involved in developing the most advanced AI models; controllability impacts the number of potential second- and third-movers who can apply AI innovations developed elsewhere; and covertness impacts the ability to monitor, measure, and assess AI capabilities and their future trajectories. AI models are often embedded in larger systems that support decision-making and include sensors, hardware components, and human-machine interfaces.[163] Future research on costly signals and AI should explore the degree to which AI-enabled systems vary in terms of costs, controllability, and covertness, as well as other technical characteristics that enable or constrain the transmission of costly AI signals.

The wide range of applications and the untested assumptions of how AI will affect crisis stability underscore both the critical need and the challenge of signaling intentions in this rapidly evolving field. Indeed, AI models and the larger systems of which they are a part complicate the task of signaling. AI models are vulnerable to intentional failures, such as the poisoning of data used to train AI models, adversarial attacks on trained AI models, and supply chain exploitation.[164] As AI-powered algorithms play a more central role in decision-making and communication, policymakers will need to grapple with the risk of AI-enabled deception, AI-driven "personalized persuasion," and unintentional signals emanating from AI agents in dynamic environments.[165] Signaling through greater transparency, information sharing, test and evaluation, and security by design across the life cycle of AI development will be critical to ensure these systems operate as intended.[166]

Policymakers should be more willing to develop and use costly signaling mechanisms with respect to AI, but they must also be aware of the limitations of this tool. Signals can be noisy, but they are an enduring feature of modern diplomacy. The answer is not to give up on the enterprise of sending costly signals, but instead to be deliberate in how and through which channels policymakers convey information in complex interdependent networks where the private sector and academic research play an important role.

One hopes that today's major powers need not experience the modern equivalent of a Cuban Missile Crisis before establishing open lines of communication and clearer understandings of the role that emerging technologies will play in crisis decision-making. The early stages of geopolitical competitions are often the most perilous for international stability. Power asymmetries loom large in the minds of policymakers, and the rules of the road are more fluid.[167] While uncertainty remains the watchword, leaders should consider the value and limitations of costly signals as a policy tool for modern AI. Talk is cheap, but inadvertent escalation is costly to all sides. By expanding the AI toolkit to include costly signals, policymakers can better communicate intent and learn from the shifting patterns of history without repeating its follies.

Authors

Andrew Imbrie is Associate Professor of the Practice in the Gracias Chair for Security and Emerging Technology at the School of Foreign Service and an Affiliate at the Center for Security and Emerging Technology at Georgetown University.

Owen J. Daniels is the Andrew W. Marshall Fellow at Georgetown's Center for Security and Emerging Technology.

Helen Toner is Director of Strategy and Foundational Research Grants at Georgetown's Center for Security and Emerging Technology and also serves in an uncompensated capacity on OpenAI's nonprofit board.

Acknowledgements

The authors are grateful to Samanvya Hooda and Jessica Maksimov for their excellent research assistance. For their insights and constructive feedback at various stages of this project, the authors would like to thank Catherine Aiken, Sam Bresnick, Margarita Konaev, Igor Mikolic-Torreira, and Emelia Probasco. Our thanks to Matt Mahoney, Jahnavi Mukul, and Shelton Fitch for editorial support. We are also indebted to Jason Brown and Dr. Erik Lin-Greenberg for their thoughtful comments and reviews. The opinions and characterizations in this piece are those of the authors and do not necessarily represent those of the U.S. government.

Document Identifier: doi: 10.51593/ 20230033

Appendix A: Multilateral examples of language about "democracy" or "democratic values" and AI

Body	Selected Document(s)	Selected Verbiage
Council of Europe	October 2020; resolution and recommendations about AI and democracy Draft convention on AI, human rights, and democracy	**From the October 2020 Resolution:** "The committee of ministers decided to give priority to…an appropriate legal framework…based on the Council of Europe's standards on human rights, democracy and the rule of law, and conducive to innovation." "There is an urgent need to set up national and international regulatory frameworks to ensure democratic governance of artificial intelligence and prevent its misuse." "the Assembly strongly believes that there is a need to create a cross-cutting regulatory framework for AI, with specific principles based on the protection of human rights, democracy and rule of law." **From the Draft Convention on AI, Human Rights, and Democracy:** Article 5: "any interference with human rights and fundamental freedoms by a public authority…. is compatible with core values of democratic societies, in accordance with the law and necessary in a democratic society in pursuit of a legitimate public interest." Article 7: "Each Party shall take all necessary measures to preserve the integrity of democratic institutions and processes…. in the context of application of an artificial intelligence system." Article 8: "…prevent and mitigate any adverse impacts of the application of an artificial intelligence system on the enjoyment of human rights and fundamental freedoms, the functioning of democracy and the observance of the rule of law in their operations."

Body	Selected Document(s)	Selected Verbiage
		Article 10: “Each Party shall take the necessary measures to ensure that all interested parties, groups and individuals enjoy equal and fair access to public debate and inclusive democratic processes.”
European Union (EU)	Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts	“The purpose of this Regulation is to promote the uptake of human centric and trustworthy artificial intelligence and to ensure a high level of protection of health, safety, fundamental rights, democracy and rule of law and the environment from harmful effects of artificial intelligence systems in the Union while supporting innovation and improving the functioning of the internal market.” “This Regulation should preserve the values of the Union facilitating the distribution of artificial intelligence benefits across society, protecting individuals, companies, democracy and rule of law and the environment from risks while boosting innovation and employment and making the Union a leader in the field.” “Certain AI systems intended for the administration of justice and democratic processes should be classified as high-risk, considering their potentially significant impact on democracy, rule of law, individual freedoms as well as the right to an effective remedy and to a fair trial.” “In order to address the risks of undue external interference to the right to vote enshrined in Article 39 of the Charter, and of disproportionate effects on democratic processes, democracy, and the rule of law, AI systems intended to be used to influence the outcome of an election or referendum or the voting behaviour of natural persons in the exercise of their vote in elections or referenda should be classified as high-risk AI systems.”

Body	Selected Document(s)	Selected Verbiage
Freedom Online Coalition	Joint Statement on AI and Human Rights	Propose ten actions items in order to "promote respect for human rights, democracy, and the rule of law in the design, development, procurement, and use of AI systems."
Global 7 (G7)	2023 Communique	**From the 2023 G7 Communique:** "Hold international discussions on inclusive artificial intelligence (AI) governance and interoperability to achieve our common vision and goal of trustworthy AI, in line with our shared democratic values."
	G7 Science and Technology Ministers' Declaration on COVID-19	**From the 2020 COVID-19 Declaration:** "...to enhance multi-stakeholder cooperation in the advancement of AI that reflects our shared democratic values."
Global Partnership on Artificial Intelligence	Launch Statement	"...we will support the responsible and human-centric development and use of AI in a manner consistent with human rights, fundamental freedoms, and our shared democratic values, as elaborated in the OECD Recommendation on AI."
Organization for Economic	Recommendation of the Council on	"...promote an AI-powered crisis response that is trustworthy and respects human-centred and democratic values."

Body	Selected Document(s)	Selected Verbiage
Cooperation and Development (OECD)	Artificial Intelligence	"AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle."
U.S.-European Union Trade and Technology Council	U.S.-EU Trade and Technology Council Inaugural Joint Statement Website	**From the Inaugural Joint Statement:** "The United States and European Union affirm their willingness and intention to develop and implement AI systems that are innovative and trustworthy and that respect universal human rights and shared democratic values." **From the U.S. Trade Representative Website:** "...cooperate on the development and deployment of new technologies based on our shared democratic values, including respect for human rights, that encourage compatible standards and regulations."
United Nations Educational, Scientific and Cultural Organization (UNESCO)	Recommendation on the Ethics of Artificial Intelligence	"...the main action is for Member States to put in place effective measures...to ensure that other stakeholders, develop human rights, rule of law, democracy, and ethical impact assessment and due diligence tools in line with guidance including the United Nations Guiding Principles on Business and Human Rights."

Body	Selected Document(s)	Selected Verbiage
Summit for Democracy	State Department Fact Sheet	"...democracies must continue *looking ahead*, so as to align emerging technologies, such as artificial intelligence (AI), with respect for democratic principles, human rights and fundamental freedoms."

Appendix B: Unilateral examples of language about "democracy" or "democratic values" and AI

Country	Selected Document(s)	Selected Verbiage
United States	Blueprint for an AI Bill of Rights; Overview	**From the Blueprint for an AI Bill of Rights and Overview Document:** "AI's important progress must not come at the price of civil rights or democratic values, foundational American principles that President Biden has affirmed as a cornerstone of his Administration." "...these (five) principles are a blueprint for building and deploying automated systems that are aligned with democratic values and protect civil rights, civil liberties, and privacy."

Country	Selected Document(s)	Selected Verbiage
	Advancing Tech for Democracy	**From Advancing Tech for Democracy:** "...development of national technology frameworks that align with human rights, and supporting the development of technologies that embed democratic values at every stage of their design and use."
United Kingdom	National AI Strategy	"...progress in AI must be achieved responsibly, according to democratic norms and the rule of law." "By leading with our democratic values, the UK will work with partners around the world to make sure international agreements embed our ethical values, making clear that progress in AI must be achieved responsibly, according to democratic norms and the rule of law."
Italy	Strategy for Technological Innovation	Italy will "engage in the promotion of an artificial intelligence that is sustainable on a social, cultural and democratic level."
Canada	Government of Canada creates Advisory Council on Artificial Intelligence	"...we can increase trust and accountability in AI while protecting our democratic values, processes and institutions."

Country	Selected Document(s)	Selected Verbiage
Brazil	Summary of Brazil AI Strategy	Strategic actions are "to stimulate actions of transparency and responsible disclosure regarding the use of AI systems, and promote the observance, by such systems, of human rights, democratic values and diversity."
Australia	Australia's Artificial Intelligence Ethics Framework	"AI systems should enable an equitable and democratic society."

Endnotes

[1] These four signaling mechanisms are drawn from and outlined in greater detail in Kai Quek, "Four Costly Signaling Mechanisms," *American Political Science Review* (2021), 115(2), 537-549, https://www.cambridge.org/core/journals/american-political-science-review/article/four-costly-signaling-mechanisms/F05A439BE1F78751453A65CADFBDB071.

[2] Evan Andrews, "Was There Really a 'Red Telephone' hotline During the Cold War," History.com, October 19, 2018, https://www.history.com/news/was-there-really-a-red-telephone-hotline-during-the-cold-war.

[3] On debates around whether to conceal or reveal military capabilities, see: Brendan Rittenhouse Green and Austin Long, "Conceal or Reveal? Managing Clandestine Military Peacetime Competition," *International Security* 44, Issue 3 (2020): 48-83, https://direct.mit.edu/isec/article-abstract/44/3/48/12283/Conceal-or-Reveal-Managing-Clandestine-Military.

[4] Sergey Radchenko and Vladislav Zubok, "Blundering on the Brink: The Secret History and Unlearned Lessons of the Cuban Missile Crisis," *Foreign Affairs*, May/June 2023, https://www.foreignaffairs.com/cuba/missile-crisis-secret-history-soviet-union-russia-ukraine-lessons.

[5] Michael Dobbs, *One Minute to Midnight: Kennedy, Khrushchev, and Castro on the Brink* (New York: Random House, 2009), 15, https://www.penguinrandomhouse.com/books/41412/one-minute-to-midnight-by-michael-dobbs/9781400078912.

[6] Raymond L. Garthoff, "The Cuban Missile Crisis: An Overview," 147, quoted in James A. Nathan, ed., *The Cuban Missile Crisis Revisited* (New York: Palgrave Macmillan, 1992).

[7] Garthoff, "The Cuban Missile Crisis," 18.

[8] Dobbs, *One Minute to* Midnight, 269.

[9] Melissa Flagg and Paul Harris, "System Re-engineering," *Center for Security and Emerging Technology*, September 2020, https://cset.georgetown.edu/publication/system-re-engineering/.

[10] Henry Farrell and Abraham L. Newman, "Weaponized Interdependence: How Global Economic Networks Shape State Coercion," *International Security* 44, Issue 1 (2019): 42-79, https://direct.mit.edu/isec/article/44/1/42/12237/Weaponized-Interdependence-How-Global-Economic.

[11] Christine H. Fox and Emelia Probasco, "Big Tech Goes to War," *Foreign Affairs*, October 19, 2022, https://www.foreignaffairs.com/ukraine/big-tech-goes-war.

[12] John Hudson and Meaghan Tobin, "Blinken Holds 'Candid' Talks with Xi Amid Effort to Ease Tensions," *Washington Post*, June 19, 2023, https://www.washingtonpost.com/national-security/2023/06/19/blinken-china-xi-jinping-meeting/.

[13] Ben Buchanan and Andrew Imbrie, *The New Fire: War, Peace, and Democracy in the Age of AI* (Cambridge: MIT Press, 2022), https://mitpress.mit.edu/9780262046541/the-new-fire/; Paul Scharre, *Four Battlegrounds: Power in the Age of Artificial Intelligence* (New York: W. W. Norton & Company, 2023), https://wwnorton.com/books/9780393866865; *Mid-Decade Challenges to National Competitiveness*, Special Competitive Studies Project, September 2022, https://www.scsp.ai/wp-content/uploads/2022/09/SCSP-Mid-Decade-Challenges-to-National-Competitiveness.pdf.

[14] Jason Matheny, "RAND President and CEO Presenting to Permanent Select Committee on Intelligence," RAND Corporation, February 28, 2023, https://www.rand.org/blog/2023/02/rand-president-and-ceo-presenting-to-house-permanent-select-committee.html.

[15] International Accounting Standard 38, Intangible Assets (IAS 38), A1492, https://www.ifrs.org/content/dam/ifrs/publications/pdf-standards/english/2021/issued/part-a/ias-38-intangible-assets.pdf.

[16] James Vincent, "OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: 'We Were Wrong,'" *The Verge*, March 15, 2023, https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview; Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," *arXiv*, February 5 2023, https://arxiv.org/abs/2302.04844.

[17] James M. Acton, "Escalation through Entanglement: How the Vulnerability of Command-and-Control Systems Raises the Risk of Inadvertent Nuclear War," *International Security* 43, Issue 1 (Summer 2018): 56-99, https://direct.mit.edu/isec/article/43/1/56/12199/Escalation-through-Entanglement-How-the; Brandi Vincent, "AI Models Will Soon Be Trained with Data From Pentagon's Technical Information Hub," *DefenseScoop*, June 8, 2023, https://defensescoop.com/2023/06/08/ai-models-will-soon-be-trained-with-data-from-pentagons-technical-information-hub/; Benjamin Jensen and Dan Tadross, "How Large-Language Models Can Revolutionize Military Planning," *War on the Rocks*, April 12, 2023, https://warontherocks.com/2023/04/how-large-language-models-can-revolutionize-military-planning/; Michael Horowitz and Erik Lin-Greenberg, "Algorithms and Influence: Artificial Intelligence and Crisis Decision, *International Studies Quarterly* 66, Issue 4 (December 2022), https://academic.oup.com/isq/article-abstract/66/4/sqac069/6753237?login=false.

[18] Heather Frase, "One Size Does Not Fit All: Assessment, Safety, and Trust for the Diverse Range of AI Products, Tools, Services, and Resources," Center for Security and Emerging Technology, February 2023, https://cset.georgetown.edu/publication/one-size-does-not-fit-all/.

[19] See, for example, James D. Fearon, "Signaling Foreign Policy Interests: Tying Hands Versus Sinking Costs," *The Journal of Conflict Resolution* 41, No. 1 (February 1997): 68-90, https://journals.sagepub.com/doi/abs/10.1177/0022002797041001004.

[20] Signals can reveal information about an actor's intentions, capabilities, resolve, goals, and time horizons. They can take the form of verbal statements or non-verbal actions and cues. Signals may be employed in coercive bargaining situations to deter an opponent or as "accommodative signals" to manage escalation risks. Decision-makers may rely on them to indicate high or low resolve, and signals can be directed toward international or domestic audiences. See, for example, Kyle Haynes, "A Question of Costliness: Time Horizons and Interstate Signaling," *Journal of Conflict Resolution* 63, Issue 8 (2019), https://journals.sagepub.com/doi/10.1177/0022002718822719; Erica D. Lonergan and Shawn W. Lonergan, "Cyber Operations, Accommodative Signaling and the De-Escalation of International Crises," *Security Studies* 31, Issue 1 (2022): 32-64, https://www.tandfonline.com/doi/abs/10.1080/09636412.2022.2040584; Amy Zegart, "Cheap Fights, Credible Threats: The Future of Armed Drones and Coercion," *Journal of Strategic Studies* 43, Issue 1 (2020): 6-46, https://www.tandfonline.com/doi/abs/10.1080/01402390.2018.1439747?casa_token=ZyJzxYp-xEAAAAAA%3AyFHoyvkqccPmd9s9aINDcsiQZ1bYia6nXt3RkPjVZRpgWo8_LmMvSuhQJppS1d7RiZbiKAeDihPj&journalCode=fjss20; Todd S. Sechser, "Reputations and Signaling in Coercive Bargaining," *Journal of Conflict Resolution* 62, Issue 2 (2018): 318-345, https://journals.sagepub.com/doi/abs/10.1177/0022002716652687; Kai Quek, "Discontinuities in Signaling Behavior Upon the Decision for War: An Analysis of China's Prewar Signaling Behavior," *International Relations of the Asia-Pacific* 15, Issue 2 (May 2015): 279-317, https://academic.oup.com/irap/article-abstract/15/2/279/735427.

[21] Zegart, "Cheap Fights, Credible Threats," 15.

[22] Samuel R. Bowman, "Eight Things to Know about Large Language Models," https://cims.nyu.edu/~sbowman/eightthings.pdf.

[23] Sebastian Rosato, *Intentions in Great Power Politics* (New Haven: Yale University Press, 2021), https://yalebooks.yale.edu/book/9780300253023/intentions-in-great-power-politics/.

[24] David M. Edelstein, *Over the Horizon: Time, Uncertainty, and the Rise of Great Powers* (Ithaca: Cornell University Press, 2020), https://www.cornellpress.cornell.edu/book/9781501748455/over-the-horizon/; David M. Edelstein, "Managing Uncertainty: Beliefs About Intentions and the Rise of Great Powers," *Security Studies* 12, Issue 1 (2006): 1-40, https://www.tandfonline.com/doi/abs/10.1080/09636410212120002.

[25] John J. Mearsheimer, *The Tragedy of Great Power Politics* (New York: W. W. Norton & Company, 2001).

[26] Robert L. Jervis, *The Logic of Images in International Relations* (New York: Columbia University Press, 1989), https://cup.columbia.edu/book/the-logic-of-images-in-international-relations/9780231069335.

[27] Jervis, *The Logic of Images in International Relations*, 18.

[28] Jervis, *The Logic of Images in International Relations*, 18.

[29] Robert F. Trager, *Diplomacy: Communication and the Origins of International Order* (Cambridge University Press, 2017), https://www.cambridge.org/us/universitypress/subjects/politics-international-relations/international-relations-and-international-organisations/diplomacy-communication-and-origins-international-order.

[30] Charles L. Glaser, "The Security Dilemma Revisited," *World Politics* 50, No. 1 (October 1997): 171-201, https://www.cambridge.org/core/journals/world-politics/article/abs/security-dilemma-revisited/0174D23352D9303257AAAC18911F3AB7.

[31] James D. Fearon, "Domestic Political Audiences and the Escalation of International Disputes," *American Political Science Review* 88, Issue 3 (1994): 577-592, https://www.cambridge.org/core/journals/american-political-science-review/article/abs/domestic-political-audiences-and-the-escalation-of-international-disputes/D22E7DE87C4CFBC436AAB3CFE7505962; Bruce Russet, Grasping the Democratic Peace (Princeton: Princeton University Press, 1993), https://press.princeton.edu/books/paperback/9780691001647/grasping-the-democratic-peace; Keren Yarhi-Milo, *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Relations* (Princeton: Princeton University Press, 2014), https://press.princeton.edu/books/hardcover/9780691159157/knowing-the-adversary; Keren Yarhi-Milo, *Who Fights for Reputation: The Psychology of Leaders in International Conflict* (Princeton: Princeton University Press, 2018), https://press.princeton.edu/books/hardcover/9780691180342/who-fights-for-reputation.

[32] Jessica L. Weeks, "Autocratic Audience Costs: Regime Type and Signaling Resolve," *International Organization* 62, Issue 1 (2008):35-64, https://www.cambridge.org/core/journals/international-organization/article/autocratic-audience-costs-regime-type-and-signaling-resolve/3EB4447D7E4584B523BFA6D5AC1542D3.

[33] H. E. Goemans, *War and Punishment: The Causes of War Termination and the First World War* (Princeton: Princeton University Press, 2000), https://press.princeton.edu/books/paperback/9780691049441/war-and-punishment.

[34] Edelstein, "Managing Uncertainty," 11.

[35] Keren Yarhi-Milo, Joshua D. Kertzer, and Jonathan Renshon, "Tying Hands, Sinking Costs, and Leader Attributes," *Journal of Conflict Resolution* 62, Issue 10, (2018): 2150-2179, https://jkertzer.sites.fas.harvard.edu/Research_files/Yarhi-Milo%20Kertzer%20Renshon%202018.pdf.

[36] These four signaling mechanisms are drawn from and outlined in greater detail in Kai Quek, "Four Costly Signaling Mechanisms," *American Political Science Review* (2021), 115(2), 537-549, https://www.cambridge.org/core/journals/american-political-science-review/article/abs/four-costly-signaling-mechanisms/F05A439BE1F78751453A65CADFBDB071. For work on sunk costs and tying hands mechanisms, see, for example, Fearon, "Signaling Foreign Policy Interests."

[37] Quek, "Four Costly Signaling Mechanisms."

[38] See, for example, "DOD Adopts Ethical Principles for Artificial Intelligence," February 24, 2020, U.S. Department of Defense, https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/; Kathleen Hicks, "What the Pentagon Thinks about Artificial Intelligence," *Politico*, June 15, 2023, https://www.politico.com/news/magazine/2023/06/15/pentagon-artificial-intelligence-china-00101751; Julian E. Barnes and David E. Sanger, "U.S. Will Try to Bring China Into Arms Control Talks," *New York Times*, June 2, 2023, https://www.nytimes.com/2023/06/02/us/politics/china-arms-control-nuclear-weapons.html?smid=nytcore-ios-share&referringSource=articleShare.

[39] "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," U.S. Department of State, February 2023, https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/.

[40] "Our Principles," Google, https://ai.google/responsibility/principles/.

[41] Jasper Hamill, "Elon Musk-founded OpenAI builds Artificial Intelligence So Powerful It Must be Kept Locked Up for the Good of Humanity," Metro, February 15, 2019, https://metro.co.uk/2019/02/15/elon-musks-openai-builds-artificial-intelligence-powerful-must-kept-locked-good-humanity-8634379/.

[42] Solaiman, "The Gradient of Generative AI Release."

[43] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, Tom Goldestein, "A Watermark for Large Language Models," *arXiv*, June 6, 2023, https://arxiv.org/pdf/2301.10226.pdf; Brandi Vincent, "Senators Propose new DOD-led Prize Competition for Tech to Detect and Watermark Generative AI," *DefenseScoop*," July 12, 2023, https://defensescoop.com/2023/07/12/senators-propose-new-dod-led-prize-competition-for-tech-to-detect-and-watermark-generative-ai/; Elias Groll, "Coming to DEF CON 31: Hacking AI Models," *CyberScoop*, May 4, 2023, https://cyberscoop.com/def-con-red-teaming-ai/?utm_source=Center+for+Security+and+Emerging+Technology&utm_campaign=03ded6ce9f-Newsletter_CAMPAIGN_2023_05_18_12_55&utm_medium=email&utm_term=0_-03ded6ce9f-%5BLIST_EMAIL_ID%5D; "Call for Research Ideas: AI Assurance for General-Purpose Systems in Open-ended Domains," Center for Security and Emerging Technology, https://cset.georgetown.edu/wp-content/uploads/FRG-Call-for-research-ideas-AI-assurance-for-general-purpose-systems-in-open-ended-domains.pdf; Rob Ashmore, Radu Calinescu, and Colin Paterson, "Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges," *arXiv*, May 10, 2019, https://arxiv.org/abs/1905.04223.

[44] Andrew Imbrie, Daniel Baer, Andrew Trask, Anna Puglisi, Erik Brattberg, and Helen Toner, "Privacy is Power: How Tech Policy Can Bolster Democracy," *Foreign Affairs*, January 19, 2022, https://www.foreignaffairs.com/articles/world/2022-01-19/privacy-power.

[45] "Emblem: Relevant Articles of the 1949 Geneva Conventions and their Additional Protocols," International Committee of the Red Cross, January 11, 2008, https://www.icrc.org/en/doc/resources/documents/misc/emblem-ihl-011108.htm.

[46] Michael Horowitz and Paul Scharre, "AI and International Stability: Risks and Confidence Building Measures," Center for a New American Security, January 12, 2021, https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures; "ICRC Position on Autonomous Weapon Systems," May 12, 2021, https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems.

[47] Written Testimony of Sam Altman, U.S. Senate Committee on the Judiciary, Subcommittee on Privacy, Technology, and the Law, May 15, 2023, https://www.judiciary.senate.gov/imo/media/doc/2023-05-16%20-%20Bio%20&%20Testimony%20-%20Altman.pdf; Matthew Mittlesteadt, "AI Verification: Mechanisms to Ensure AI Arms Control Compliance," Center for Security and Emerging Technology, February 2021, https://cset.georgetown.edu/publication/ai-verification/.

[48] "Dos, Don'ts and Geo-fencing: Europe Proposes Rules for Small Drones," *Reuters*, May 5, 2017, https://www.reuters.com/article/us-europe-drones-idUSKBN1811KD.

[49] Quek, "Four Costly Signaling Mechanisms," 539.

[50] Quek, "Four Costly Signaling Mechanisms," 539.

[51] Mittlesteadt, "AI Verification."

[52] Yonadav Shavit, "What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Training via Compute Monitoring," arXiv, March 20, 2023, https://arxiv.org/abs/2303.11341; Krystal Jackson, Karson Elmgren, Jacob Feldgoise, and Andrew Critch, "Compute Accounting Principles Can Help Reduce AI Risk." Tech Policy Press, November 30, 2022, https://techpolicy.press/compute-accounting-principles-can-help-reduce-ai-risks/; Jason Matheny, "Advancing Trustworthy Artificial Intelligence," Testimony presented before the U.S. House Committee on Science, Space and Technology, June 22, 2023, https://republicans-science.house.gov/_cache/files/4/f/4f36e843-0ef0-4964-b283-4c165bac6250/ACFFA7F14FA3397CD845977D632D307A.2023-06-22-dr.-matheny-testimony.pdf; Jason Matheny, "Here's a Simple Way to Regulate Powerful AI Models," *Washington Post*, August 16, 2023, https://www.washingtonpost.com/opinions/2023/08/16/ai-danger-regulation-united-states/.

[53] Frase, "One Size Does Not Fit All"; Dewey Murdick, Testimony presented before the U.S. House Committee on Science, Space and Technology," June 22, 2023, https://republicans-science.house.gov/_cache/files/b/1/b120b2e0-d60f-4aa1-a9e6-f03ae7c4c718/C478786EEE329201D106582854D95DCF.2023-06-22-dr.-murdick-testimony.pdf.

[54] Sasha Costanza-Chock, Joy Buolamwini, Inioluwa Deborah Raji, "Who Audits the Auditors? Recommends. From a Field Scan of the Algorithmic Auditing Ecosystem," ACM Conference on Fairness, Accountability, and Transparency (FAccT), https://www.ajl.org/auditors.

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

[56] Zachary Arnold and Helen Toner, "AI Accidents: An Emerging Threat," *Center for Security and Emerging Technology*, July 2021, https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/.

[57] Thomas J. Miceli and Kathleen Segerson, "Liability Versus Regulation for Dangerous Products When Consumers Vary in Their Susceptibility to Harm and May Misperceive Risk," *Review of Law & Economics* 9, No. 3 (2013): 341-355, https://www.degruyter.com/document/doi/10.1515/rle-2013-0004/html; Jack Clark, Kyle Miller, and Rebecca Gelles, "Measuring AI Development: A Prototype Methodology to Inform Policy," December 2021, Center for Security and Emerging Technology, https://cset.georgetown.edu/publication/measuring-ai-development/; Jess Whittlestone and Jack Clark, "Why and How Governments Should Monitor AI Development," *arXiv*, August 31, 2023, https://arxiv.org/abs/2108.12427.

[58] Quek, "Four Costly Signaling Mechanisms," 540-542.

[59] See, for example, "The New START Treaty: Central Limits and Key Provisions," Congressional Research Service, February 2, 2022, https://sgp.fas.org/crs/nuke/R41219.pdf.

[60] Margaret Mitchell et al., "Model Cards for Model Reporting," *arXiv*, January 14, 2019, https://arxiv.org/abs/1810.03993; Brad Smith, "How Do We Best Govern AI," Microsoft (blog), May 25, 2023, https://blogs.microsoft.com/on-the-issues/2023/05/25/how-do-we-best-govern-ai/.

[61] "OECD AI Principles Overview," OECD Artificial Intelligence Policy Observatory, https://oecd.ai/en/ai-principles.

[62] Matheny, Testimony presented before the U.S. House Committee on Science, Space and Technology; Rumman Chowdhury and Jutta Williams, "Introducing Twitter's First Algorithmic Bias Bounty Challenge," Twitter (blog), July 20, 2021, https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge.

[63] Daniel Zhang et al., "Enhancing International Cooperation in AI Research: The Case for a Multilateral AI Research Institute," Stanford University Human-Centered Artificial Intelligence, May 2022, https://hai.stanford.edu/white-paper-enhancing-international-cooperation-ai-research-case-multilateral-ai-research-institute; Sam Altman, Greg Brockman, Ilya Sutskever, "Governance of Superintelligence," OpenAI, May 22, 2023, https://openai.com/blog/governance-of-superintelligence.

[64] Charlotte Stix, "Foundations for the Future: Institution Building For the Purpose of Artificial Intelligence Governance," *AI and Ethics* 2 (2022): 463-476, https://link.springer.com/article/10.1007/s43681-021-00093-w.

[65] The authors are grateful to Dr. Erik Lin-Greenberg for his insights on this topic.

[66] Tim G. J. Rudner and Helen Toner, "Key Concepts in AI Safety: Interpretability in Machine Learning," Center for Security and Emerging Technology, March 2021, https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-interpretability-in-machine-learning/; Timothy B. Lee and Sean Trott, "Large Language Models, Explained with a Minimum of Math and Jargon," Understanding AI, July 27, 2023, https://www.understandingai.org/p/large-language-models-explained-with. Similar challenges may arise with efforts to label, or "watermark," AI text and with techniques for aligning AI models with human objectives. See, for example, Vinu Sankar Sadasivan, "Can AI-Generated Text Be Reliably Detected," *arXiv*, March 17, 2023, https://arxiv.org/abs/2303.11156; Stephen Casper et al., "Open Problems and Fundamental Limitations of Reinforcement Learning From Human Feedback," *arXiv*, July 27, 2023, https://arxiv.org/abs/2307.15217.

[67] The authors are grateful to Dr. Erik Lin-Greenberg for highlighting the complexities of intentional and unintentional signaling in the context of AI.

[68] Amandeep Singh Gill, "The Role of the United Nations in Addressing Emerging Technologies in the Area of Lethal Autonomous Weapons Systems," *UN Chronicle*, December 2018, https://www.un.org/en/un-chronicle/role-united-nations-addressing-emerging-technologies-area-lethal-autonomous-weapons.

[69] Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, September 25, 2019, https://documents.unoda.org/wp-content/uploads/2020/09/CCW_GGE.1_2019_3_E.pdf.

[70] Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 13.

[71] "The poistion (sic) paper submitted by the Chinese delegation to CCW 5th Review Conference," last accessed: August 25, 2023, https://perma.cc/YEG4-3GJZ.

[72] "The poistion (sic) paper submitted by the Chinese delegation to CCW 5th Review Conference," last accessed: August 25, 2023, https://perma.cc/YEG4-3GJZ; Elsa Kania, "China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems," *Lawfare*, April 17, 2018, https://www.lawfaremedia.org/article/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems.

[73] Position Paper Submitted by China, "Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects," April 9-13, 2018, https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/documents/GGE.1-WP7.pdf.

[74] Position Paper Submitted by China, April 9-13, 2018.

[75] Kania, "China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems."

[76] Peter Mattis, "China's 'Three Warfares' in Perspective," *War on the Rocks*, January 30, 2018, https://warontherocks.com/2018/01/chinas-three-warfares-perspective/; Buchanan and Imbrie, *The New Fire*, 150.

[77] Ryan Fedasiuk, Jennifer Melot, and Ben Murphy, "Harnessed Lightning: How the Chinese Military is Adopting Artificial Intelligence," Center for Security and Emerging Technology, October 2021, 7-13, 47-57, https://cset.georgetown.edu/publication/harnessed-lightning/.

[78] Gregory C. Allen, Testimony Before the U.S.-China Economic and Security Review Commission, "China's Pursuit of Defense Technologies: Implications for U.S. Multilateral Export Control and Investment Screening Regimes," Panel II: Obstacles and Breakthroughs in China's Defense Technological Development, April 13, 2023, https://www.uscc.gov/sites/default/files/2023-04/Gregory_Allen_Testimony.pdf. See also: Elsa B. Kania, "'AI Weapons' in China's Military Innovation," Brookings Institution, April 2020, https://www.brookings.edu/articles/ai-weapons-in-chinas-military-innovation/; Fedasiuk, Melot, and Murphy, "Harnessed Lightning."

[79] Allen, Testimony Before the U.S.-China Economic and Security Review Commission, 7. As Allen notes, "I transcribed Zeng's comments (as provided by the simultaneous translators) as I was in attendance at the same conference. However, in the subsequently released transcript of the conference session, all mention of Zeng's presentation and participation was removed, likely indicating that the Chinese government censors had determined it was not in China's interest to have that information in the open."

[80] Allen, Testimony Before the U.S.-China Economic and Security Review Commission, 7.

[81] "DoD Announces Update to DoD Directive 3000.09 'Autonomy in Weapon Systems'", U.S. Department of Defense, January 25, 2023, https://www.defense.gov/News/Releases/Release/Article/3278076/dod-announces-update-to-dod-directive-300009-autonomy-in-weapon-systems/.

[82] Arnold and Toner, "AI Accidents."

[83] Michele A. Flournoy, Avril Haines, and Gabrielle Chefitz, "Building Trust through Testing," *Center for Security and Emerging Technology*, October 2020, https://cset.georgetown.edu/event/building-trust-through-testing/.

[84] Flournoy, Haines, and Chefitz, "Building Trust through Testing"; Heather M. Roff and David Danks, "'Trust but Verify'" The Difficulty of Trusting Autonomous Weapons Systems," *Journal of Military Ethics* 17 (1) (2018): 2-20, https://philpapers.org/rec/ROFTBV.

[85] Tai Ming Cheung, *Innovate to Dominate: The Rise of the Chinese Techno-Security State* (Ithaca: Cornell University Press, 2022), https://www.cornellpress.cornell.edu/book/9781501764349/innovate-to-dominate/#bookTabs=1; Ryan Fedasiuk, Karson Elmgren, and Ellen Lu, "Silicon Twist: Managing the Chinese Military's Access to AI Chips," Center for Security and Emerging Technology, June 2022, https://cset.georgetown.edu/publication/silicon-twist/; Elsa B. Kania and Lorand Laskai, "Myths and Realities of China's Military-Civil Fusion Strategy," *Center for a New American Security*, January 28,

2021, https://www.cnas.org/publications/reports/myths-and-realities-of-chinas-military-civil-fusion-strategy.

[86] "Commerce Implements New Export Controls on Advanced Computing and Semiconductor Manufacturing Items to the People's Republic of China (PRC)," Bureau of Industry and Security, U.S. Department of Commerce, October 7, 2022, https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3158-2022-10-07-bis-press-release-advanced-computing-and-semiconductor-manufacturing-controls-final/file.

[87] Cade Metz, "Away from Silicon Valley, the Military Is the Ideal Customer," *New York Times*, February 26, 2021, https://www.nytimes.com/2021/02/26/technology/anduril-military-palmer-luckey.html.

[88] Will Knight, "Russia's Killer Drone in Ukraine Raises Fears About AI in Warfare," *Wired*, March 17, 2022, https://www.wired.com/story/ai-drones-russia-ukraine/.

[89] Gregory C. Allen, "Russia Probably Has Not Used AI-Enabled Weapons in Ukraine, but That Could Change," Center for Strategic and International Studies, May 26, 2022, https://www.csis.org/analysis/russia-probably-has-not-used-ai-enabled-weapons-ukraine-could-change.

[90] Allen, "Russia Probably Has Not Used AI-Enabled Weapons in Ukraine, but That Could Change."

[91] Paul Scharre, "Is the UK Brimstone an 'Autonomous Weapon'?" Twitter, May 27, 2023, https://twitter.com/paul_scharre/status/1662511490925051904. See also: Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: W. W. Norton & Company, 2018).

[92] Scharre, "Is the UK Brimstone an 'Autonomous Weapon'?"

[93] Sydney J. Freedberg Jr., "No AI for Nuclear Command & Control: JAIC's Shanahan," *Breaking Defense*, September 25, 2019, https://breakingdefense.com/2019/09/no-ai-for-nuclear-command-control-jaics-shanahan/. See also: "2022 National Defense Strategy of the United States of America: Including the 2022 Nuclear Posture Review the 2022 Missile Defense Review, 13, https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.PDF#page=33.

[94] "Building Consensus on the U.S. Framework for a Political Declaration on the Responsible Military Use of Artificial Intelligence and Autonomy," U.S. Department of State, February 16, 2023, https://www.state.gov/building-consensus-on-the-u-s-framework-for-a-political-declaration-on-the-responsible-military-use-of-artificial-intelligence-and-autonomy/.

[95] Amanda Macias, "Russia's Nuclear-Armed Underwater Drone May Be Ready for War in Eight Years," March 25, 2019, CNBC, https://www.cnbc.com/2019/03/25/russias-nuclear-armed-underwater-drone-may-be-ready-for-war-in-2027.html; Silky Kaur, "One Nuclear-armed Poseidon Torpedo Could Decimate a Coastal City. Russia Wants 30 of Them," *Bulletin of the Atomic Scientists*, June 14, 2023, https://thebulletin.org/2023/06/one-nuclear-armed-poseidon-torpedo-could-decimate-a-coastal-city-russia-wants-30-of-them/.

[96] Tony Munroe, Andrew Osborn, and Humeyra Pamuk, "China, Russia Partner Up Against West at Olympics Summit," *Yahoo News*, February 4, 2022, https://news.yahoo.com/russia-china-tell-nato-stop-101435894.html.

[97] Andrew Imbrie and Elsa B. Kania, "AI Safety, Security, and Stability among the Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement," *Center for Security and Emerging Technology*, December 2019, https://cset.georgetown.edu/publication/ai-safety-security-and-stability-among-great-powers-options-challenges-and-lessons-learned-for-pragmatic-engagement/; Fu Ying and John Allen, "Together, The U.S. And China Can Reduce The Risks From AI," *Noema*, December 17, 2020, https://www.noemamag.com/together-the-u-s-and-china-can-reduce-the-risks-from-ai/; John R. Allen, Ryan Hass, and Bruce Jones, "Rising to the Challenge: Navigating Competition, Avoiding Crisis, and Advancing U.S. Interests in Relations With China," Brookings Institution, November 2021, https://www.brookings.edu/wp-content/uploads/2021/11/FP_20211105_us_china_rivalry_hass_jones_allen.pdf.

[98] Allen, Testimony Before the U.S.-China Economic and Security Review Commission, April 13, 2023.

[99] See, for example, A. Feder Cooper, Karen Levy, and Christopher De Sa, "Accuracy-Efficiency Trade-Offs and Accountability in Distributed ML Systems," *arXiv*, October 2,2 021, https://arxiv.org/pdf/2007.02203.pdf.

[100] Imbrie and Kania, "AI Safety, Security, and Stability among the Great Powers."

[101] Buchanan and Imbrie, *The New Fire*, 211-230.

[102] Caitlin Talmadge, "Would China Go Nuclear? Assessing the Risk of Chinese Nuclear Escalation in a Conventional War with the United States," *International Security* 41, Issue 4 (2017): 50-92, https://www.belfercenter.org/publication/would-china-go-nuclear-assessing-risk-chinese-nuclear-escalation-conventional-war; Horowitz and Erik Lin-Greenberg, "Algorithms and Influence."

[103] "Agreement Between the Government of the United States of America the Government of the Union of Soviet Socialist Republics on the Prevention of Incidents On and Over the High Seas," U.S. Department of State, https://2009-2017.state.gov/t/isn/4791.htm.

[104] Horowitz and Scharre, "AI and International Stability: Risks and Confidence Building Measures," 16.

[105] Mittlesteadt, "AI Verification."

[106] Husanjot Chahal, Helen Toner, and Ilya Rahkovsky, "Small Data's Big AI Potential," Center for Security and Emerging Technology, September 2021, https://cset.georgetown.edu/publication/small-datas-big-ai-potential/.

[107] Paul Mozur, "One Month, 500,000 Face Scans: How China is Using A.I. to Profile a Minority," *The New York Times*, April 14, 2019, https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html; Jane Wakefield, "AI Emotion-Detection Software Tested on Uyghurs," *British Broadcasting Corporation*, 26 May 2021, https://www.bbc.com/news/technology-57101248; Avril Haines, "Countering the Misuse of Technology and the Rise of Digital Authoritarianism," Address to the Summit for Democracy 2023, March 30, 2023, https://www.dni.gov/index.php/newsroom/speeches-interviews/speeches-interviews-2023/item/2374-summit-for-democracy-2023.

[108] "G7 Hiroshima Leaders' Communiqué," Group of 7, Hiroshima, May 20, 2023, https://www.mofa.go.jp/files/100506878.pdf.

[109] "Artificial Intelligence Act Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))," European Parliament, June 23, 2030, 2, https://artificialintelligenceact.eu/wp-content/uploads/2023/06/AIA-%E2%80%93-IMCO-LIBE-Draft-Compromise-Amendments-14-June-2023.pdf. A number of amendments adopted by the European Parliament also added language emphasizing AI's impact on democracy and the rule of law. Examples may be found on pages 3, 10, 19, 23, 36, and 55, among others.

[110] "OECD AI Principles Overview," OECD.AI Policy Observatory, https://oecd.ai/en/ai-principles; "Artificial Intelligence: Ensuring Respect for Democracy, Human Rights and the Rule of Law," Council of Europe, https://pace.coe.int/en/pages/artificial-intelligence; "Revised Zero Draft [Framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law," Committee on Artificial Intelligence, Strasbourg, Council of Europe, January 6, 2023, https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193f; "Working Group on Responsible AI," Global Partnership on Artificial Intelligence, https://gpai.ai/projects/responsible-ai/; "Recommendation on the

Ethics of Artificial Intelligence," United Nations Educational, Scientific and Cultural Organization, 2021, https://unesdoc.unesco.org/ark:/48223/pf0000380455; "FOC Joint Statement on Artificial Intelligence and Human Rights," Freedom Online Coalition, November, 2020, https://freedomonlinecoalition.com/wp-content/uploads/2021/06/FOC-Joint-Statement-on-Artificial-Intelligence-and-Human-Rights.pdf; "U.S.-EU Trade and Technology Council Inaugural Joint Statement," The White House, September 29, 2021, https://www.whitehouse.gov/briefing-room/statements-releases/2021/09/29/u-s-eu-trade-and-technology-council-inaugural-joint-statement/; "Proposal for a Regulation of the European Parliament and of the Council on Harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts," Committee on the Internal Market and Consumer Protection Committee on Civil Liberties, Justice and Home Affairs, European Parliament, May 9, 2023, https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf.

[111] "Australia's AI Ethics Principles," Department of Industry, Science and Resources, Government of Australia, https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles; "Summary of the Brazilian Artificial Intelligence Strategy," Ministry of Science, Technology and Innovations, Government of Brazil, 2021, https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/ebia-summary_brazilian_4-979_2021.pdf; "Government of Canada Creates Advisory Council on Artificial Intelligence," Innovation, Science and Economic Development Canada, Ottawa, Ontario: Government of Canada, May 14, 2019, https://www.canada.ca/en/innovation-science-economic-development/news/2019/05/government-of-canada-creates-advisory-council-on-artificial-intelligence.html; "Strategia Per L'Innovazione Tecnologica e La Digitalizzazione del Paese," Rome, Government of Italy, https://assets.innovazione.gov.it/1610546390-midbook2025.pdf; "Government Use of Artificial Intelligence in New Zealand," The Law Foundation and University of Otago, 2019, https://www.data.govt.nz/assets/data-ethics/algorithm/NZLF-report.pdf; "Estrategia Nacional de Inteligencia Artificial," Ministry of Economic Affairs and Digital Transformation, Madrid, Government of Spain, November 2020, https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/021220-ENIA.pdf; "National AI Strategy," Office for Artificial Intelligence, London, Government of the United Kingdom, 2022, https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version.

[112] "Blueprint for an AI Bill of Rights," Office of Science and Technology Policy, Washington, D.C., The White House, https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[113] "France AI Strategy Report," European Commission, Brussels, European Union, https://ai-watch.ec.europa.eu/countries/france/france-ai-strategy-report_en#ai-to-address-societal-challenges; "Artificial Intelligence Strategy of the German Federal Government," Berlin, Government of Germany, December, 2020, https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-

Strategie_engl.pdf; "AI Strategy 2022," Secretariat of Science, Technology and Innovation Policy Cabinet office, Tokyo, Government of Japan, August, 2022, https://www8.cao.go.jp/cstp/ai/aistratagy2022en_ov.pdf; "National Strategy for Artificial Intelligence," Ministry of Science and ICT, Seoul, Government of South Korea, https://www.msit.go.kr/bbs/view.do?sCode=eng&nttSeqNo=9&bbsSeqNo=46&mId=10&mPid=9. Germany removed several mentions of democracy and democratic values from the 2020 version of its AI strategy that were present in the 2018 iteration of the strategy. Meanwhile, despite France's AI document not explicitly mentioning democracy, French President Emmanuel Macron acknowledged AI's impact on democracy in an interview, stating, "Europe is the place where the DNA of democracy was shaped, and therefore I think Europe has to get to grips with [AI that] could become a big challenge for democracies." See: "Artificial Intelligence Strategy of the German Federal Government," Berlin: Government of Germany, December 2020, https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf; "Artificial Intelligence Strategy," Berlin, Government of Germany, November 2018, https://www.ki-strategie-deutschland.de/home.html?file=files/downloads/Nationale_KI-Strategie_engl.pdf; Nicholas Thompson, "Emmanuel Macron Talks to WIRED About France's AI Strategy," *Wired*, March 31, 2018, https://www.wired.com/story/emmanuel-macron-talks-to-wired-about-frances-ai-strategy/.

[114] Summit for Democracy, U.S. Department of State, March 2023, https://www.state.gov/summit-for-democracy-2023/; "Fact Sheet: Advancing Technology for Democracy," The White House, March 29, 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/03/29/fact-sheet-advancing-technology-for-democracy-at-home-and-abroad/.

[115] "At Summit for Democracy, the United States and the United Kingdom Announce Winners of Challenge to Drive Innovation in Privacy-Enhancing Technologies That Reinforce Democratic Values," Office of Science and Technology Policy, The White House, March 31, 2023, https://www.whitehouse.gov/ostp/news-updates/2023/03/31/us-uk-annouce-winners-innovation-pets-democratic-values/; "Fact Sheet: Announcing the Presidential Initiative for Democratic Renewal," The White House, December 9, 2021, https://www.whitehouse.gov/briefing-room/statements-releases/2021/12/09/fact-sheet-announcing-the-presidential-initiative-for-democratic-renewal/.

[116] "Framework for G7 Collaboration on Digital Technical Standards, United Kingdom," Group of 7, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/986159/Annex_1__Framework_for_G7_collaboration_on_Digital_Technical_Standards.pdf.

[117] "Stop Killer Robots," https://www.stopkillerrobots.org/; "How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net: Questions and Answers," Human Rights Watch, November

10, 2021, https://www.hrw.org/news/2021/11/10/how-eus-flawed-artificial-intelligence-regulation-endangers-social-safety-net.

[118] Summit for Democracy, "Call to the Private Sector to Advance Democracy," https://www.state.gov/wp-content/uploads/2023/02/Private-Sector-Call-to-Advance-Democracy-1.pdf.

[119] Raphael Koster et al., "Human-Centered Mechanism Design with Democratic AI," Google DeepMind, July 4, 2022, https://www.deepmind.com/blog/human-centred-mechanism-design-with-democratic-ai.

[120] Raphael Koster et al., "Human-Centered Mechanism Design with Democratic AI," *Nature Human Behaviour* Volume 6, p. 1398–1407, https://www.nature.com/articles/s41562-022-01383-x.

[121] Mark Hosenball, "Factbox: Key findings from Senate inquiry into Russian interference in 2016 U.S. election," Reuters, August 18, 2020, https://www.reuters.com/article/us-usa-trump-russia-senate-findings-fact/factbox-key-findings-from-senate-inquiry-into-russian-interference-in-2016-u-s-election-idUSKCN25E2OY; Ninon Bulckaert, "How France Successfully Countered Russian Interference During the Presidential Election," *Euractiv*, July 17, 2018, https://www.euractiv.com/section/elections/news/how-france-successfully-countered-russian-interference-during-the-presidential-election/.

[122] Josh A. Goldstein, Girish Sastry, Micah Musser, et al., "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations," Center for Security and Emerging Technology, OpenAI, and Stanford Internet Observatory, January, 2023, https://arxiv.org/pdf/2301.04246.pdf.

[123] "Statement by Vice President Joe Biden on Foreign Interference in U.S. Elections," The American Presidency Project, July 20, 2020, https://www.presidency.ucsb.edu/documents/statement-vice-president-joe-biden-foreign-interference-us-elections.

[124] Colin Clark, "SecDef Austin Commits US to 'Responsible AI,'" *Breaking Defense,* July 13, 2021, https://breakingdefense.com/2021/07/secdef-austin-commits-us-to-responsible-ai/.

[125] Jared Cohen, "The Rise of Geopolitical Swing States," Goldman Sachs, May 15, 2023, https://www.goldmansachs.com/intelligence/pages/the-rise-of-geopolitical-swing-states.html; Heather Conley et al., "Alliances in a Shifting Global Order: Rethinking Transatlantic Engagement with Global Swing States," German Marshall Fund of the United States, 2023, https://www.gmfus.org/sites/default/files/2023-04/Global%20Swing%20States_27%20apr_FINAL_embargoed%20until%202%20May%202023.pdf.

[126] Sheena Chestnut Greitens, "Dealing with Demand for China's Global Surveillance Exports," Brookings Institution, April, 2020, https://www.brookings.edu/wp-content/uploads/2020/04/FP_20200428_china_surveillance_greitens_v3.pdf.

[127] U.S.-Gulf ties, most notably to Saudi Arabia, date to as far back as World War II. F. Gregory Gause, "British and American Policies in the Persian Gulf, 1968-1973," *Review of International Studies* 11, no. 4 (1985): 247–73. http://www.jstor.org/stable/20097054; Mehran Kamrava, *Troubled Waters: Insecurity in the Persian Gulf* (Ithaca: Cornell University Press, 2018), 1-32; Dennis Ross, "Considering Soviet Threats to the Persian Gulf," *International Security* 6, Issue 2 (1981): 159–80, https://doi.org/10.2307/2538650; Geoffrey F. Gresh, *Gulf Security and the U.S. Military: Regime Survival and the Politics of Basing* (Stanford: Stanford University Press, 2015), 19-73. https://doi.org/10.2307/j.ctvqsdsj6.

[128] *The Military Balance 2023*, International Institute for Strategic Studies, 2023, https://www.iiss.org/publications/the-military-balance/; "Qatar: Issues for the 118th Congress," Congressional Research Service, June 21, 2023, https://sgp.fas.org/crs/mideast/R47467.pdf; U.S. Naval Forces Central Command, https://www.cusnc.navy.mil/; Combined Maritime Forces, U.S. Naval Forces Central Command, https://www.cusnc.navy.mil/Combined-Maritime-Forces/; "United States Central Command," Congressional Research Service, December 16, 2022, https://crsreports.congress.gov/product/pdf/IF/IF11428; Matthew Wallin, "U.S. Military Bases and Facilities in the Middle East," American Security Project, June, 2018, https://www.americansecurityproject.org/wp-content/uploads/2018/06/Ref-0213-US-Military-Bases-and-Facilities-Middle-East.pdf.

[129] Tobias Borck, "The Gulf States and the Iran Nuclear Deal: Between a Rock and a Hard Place," The Royal United Services Institute for Defence and Security Studies, November 29, 2021, https://rusi.org/explore-our-research/publications/commentary/gulf-states-and-iran-nuclear-deal-between-rock-and-hard-place; Eva Thiébaud, "UAE's High-Tech Toolkit for Mass Surveillance and Repression," *Le Monde Diplomatique*, January, 2023, https://mondediplo.com/2023/01/05uae; "Saudi Arabia Codifies Male Guardianship and Gender Discrimination," Amnesty International, December 9, 2022, https://www.amnesty.org/en/latest/research/2022/12/saudi-arabia-codifies-male-guardianship-and-gender-discrimination/; "Saudi Arabia: Official Hate Speech Targets Minorities," Human Rights Watch, September 26, 2017, https://www.hrw.org/news/2017/09/26/saudi-arabia-official-hate-speech-targets-minorities; "Reality Check: Migrant Workers Rights with Two Years to Qatar 2022 World Cup," Amnesty International, 2020, https://www.amnesty.org/en/latest/campaigns/2019/02/reality-check-migrant-workers-rights-with-two-years-to-qatar-2022-world-cup/; Shane Harris, Greg Miller and Josh Dawsey, "CIA Concludes Saudi Crown Prince Ordered Jamal Khashoggi's Assassination," *The Washington Post*, November 16, 2018, https://www.washingtonpost.com/world/national-security/cia-concludes-saudi-crown-prince-ordered-jamal-khashoggis-assassination/2018/11/16/98c89fe6-e9b2-11e8-a939-9469f1166f9d_story.html; Government Accountability Office, "Yemen: State and DOD Need

Better Information on Civilian Impacts of U.S. Military Support to Saudi Arabia and the United Arab Emirates," June 15, 2022, https://www.gao.gov/products/gao-22-105988; Chris Murphy, "It's Time to Rethink the U.S.-Saudi Relationship," *Foreign Policy,* July 13, 2022, https://foreignpolicy.com/2022/07/13/biden-saudi-arabia-middle-east-trip-oil-khashoggi-yemen-human-rights/.

[130] Anoushiravan Ehteshami, "China's Grand Vision and the Persian Gulf," Istituto Affari Internazionali, March 7, 2023, https://www.iai.it/sites/default/files/iaip2307.pdf.

[131] Thomas Carothers, Benjamin Press, "Security Dilemma in U.S. Foreign Policy: Lessons from Egypt, India and Turkey," Carnegie Endowment for International Peace, November 4, 2021, https://carnegieendowment.org/2021/11/04/navigating-democracy-security-dilemma-in-u.s.-foreign-policy-lessons-from-egypt-india-and-turkey-pub-85701.

[132] James Lynch, "Iron Net: Digital Repression in the Middle East and North Africa," European Council on Foreign Relations, June 29, 2022, https://ecfr.eu/publication/iron-net-digital-repression-in-the-middle-east-and-north-africa/; Sanam Vakil, Understanding the GCC Collective Security Mindset, Chatham House, November 30, 2022, https://kalam.chathamhouse.org/articles/understanding-the-gcc-collective-security-mindset; Paul Mozur and Adam Satariano, "A.I., Brain Scans and Cameras: The Spread of Police Surveillance Tech," *The New York Times,* March 30, 2023, https://www.nytimes.com/2023/03/30/technology/police-surveillance-tech-dubai.html.

[133] Ashley Roque, "As US Worries Over Beijing's Influence in Middle East, Chinese Defense Firms Glock to IDEX 2023," *Breaking Defense,* February 22, 2023, https://breakingdefense.com/2023/02/as-us-worries-over-beijings-influence-in-middle-east-chinese-defense-firms-flock-to-idex-2023/.

[134] Alexander Cornwell, "U.S. Flags Huawei 5G network Security Concerns to Gulf Allies," Reuters, September 12, 2019, https://www.reuters.com/article/us-huawei-security-usa-gulf/u-s-flags-huawei-5g-network-security-concerns-to-gulf-allies-idUSKCN1VX241.

[135] Sophie Zinser, "China's Digital Silk Road Grows with 5G in the Middle East," *The Diplomat*, December 16, 2020, https://thediplomat.com/2020/12/chinas-digital-silk-road-grows-with-5g-in-the-middle-east/.

[136] Mohammed Soliman, "The Gulf has a 5G Conundrum and Open RAN is the Key to its Tech Sovereignty," Middle East Institute, January 12, 2022, https://www.mei.edu/publications/gulf-has-5g-conundrum-and-open-ran-key-its-tech-sovereignty; Martijn Rasser and Ainikki Riikonen, "Open Future: The Way Forward on 5G," Center for a New American Security, July 28, 2020, https://www.cnas.org/publications/reports/open-future.

[137] Laurens Cerulus and Sarah Wheaton, "How Washington Chased Huawei Out of Europe," *Politico*, November 23, 2022, https://www.politico.eu/article/us-china-huawei-europe-market/.

[138] Gordon Lubold and Warren P. Strobel, "Secret Chinese Port Project in Persian Gulf Rattles U.S. Relations With U.A.E.," *The Wall Street Journal*, November 19, 2021, https://www.wsj.com/amp/articles/us-china-uae-military-11637274224; Grant Rumley, "Unpacking the UAE F-35 Negotiations," Washington Institute for Near East Policy, February 15, 2022, https://www.washingtoninstitute.org/policy-analysis/unpacking-uae-f-35-negotiations; Ashley Roque, "Potential F-35, Reaper Deal with UAE Not Completely Dead, Senior US Official Says," *Breaking Defense*, February 22, 2023, https://breakingdefense.com/2023/02/potential-f-35-reaper-deal-with-uae-not-completely-dead-senior-us-official-says/.

[139] Anna Gross, Madhumita Murgia, and Yuan Yang, "Chinese Tech Groups Shaping UN Facial Recognition Standards," *Financial Times,* December 1, 2019, https://www.ft.com/content/c3555a3c-0d3e-11ea-b2d6-9bf4d1957a67.

[140] Kayla Goode, Heeu Millie Kim, and Melissa Deng, "Examining Singapore's AI Progress," Center for Security and Emerging Technology, March 2023, https://cset.georgetown.edu/publication/examining-singapores-ai-progress/.

[141] Husanjot Chahal, Sara Abdulla, Jonathan Murdick, and Ilya Rahkovsky, "Mapping India's AI Potential," Center for Security and Emerging Technology, March 2021, https://cset.georgetown.edu/publication/mapping-indias-ai-potential/; Husanjot Chahal, Ngor Luong, Sara Abdulla, and Margarita Konaev, "Quad AI: Assessing AI-related Collaboration between the United States, Australia, India, and Japan," Center for Security and Emerging Technology, May 2022, https://cset.georgetown.edu/publication/quad-ai/.

[142] Carothers, Press, "Security Dilemma in U.S. Foreign Policy"; Jon Alterman, "U.S. Power and Influence in the Middle East: Part One," Center for Strategic and International Studies, March 7, 2022, https://www.csis.org/analysis/us-power-and-influence-middle-east-part-one.

[143] Various Authors, "The Nonaligned World," *Foreign Affairs,* May/June 2023, https://www.foreignaffairs.com/content-packages/nonaligned-world.

[144] Christine H. Fox and Emelia Probasco, "Volunteer Force: U.S. Tech Companies and Their Contributions in Ukraine," Center for Security and Emerging Technology, May 2023, https://cset.georgetown.edu/publication/volunteer-force/.

[145] Dylan Patel and Afzal Ahmad, "Google 'We Have No Moat, and Neither Does OpenAI," *semianalysis*, May 4, 2023, https://www.semianalysis.com/p/google-we-have-no-moat-and-neither.

[146] Smith, "How Do We Best Govern AI?"; "Google C.E.O. Sundar Pichai on Bard, AI 'Whiplash' and Competing with ChatGPT," New York Times Hard Fork, March 31, 2023, https://www.nytimes.com/2023/03/31/podcasts/hard-fork-sundar.html?showTranscript=1.

[147] On different approaches to release policies and the risks of LLMs leaking, see James Vincent, "Meta's Powerful AI Language Models Has Leaked Online—What Happens Now? *The Verge*, March 8, 2023, https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse.

[148] "GPT-4 System Card," OpenAI, March 23, 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[149] Nitasha Tiku, Gerrit De Vynck, and Will Oremus, "Big Tech Was Moving Cautiously on AI. Then Came ChatGPT," *Washington Post*, February 3, 2023, https://www.washingtonpost.com/technology/2023/01/27/chatgpt-google-meta/.

[150] Nico Grant, "Google Calls In Help From Larry Page and Sergey Brin for A.I. Fight," *New York Times*, February 23, 2023, https://www.nytimes.com/2023/01/20/technology/google-chatgpt-artificial-intelligence.html.

[151] Gerrit De Vynck, "ChatGPT Maker OpenAI Faces A Lawsuit Over How It Used People's Data," *Washington Post*, June 28, 2023, https://www.washingtonpost.com/technology/2023/06/28/openai-chatgpt-lawsuit-class-action/; Billy Perrigo, Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic," TIME, January 18, 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/; Matt Burgess, "The Hacking of ChatGPT Is Just Getting Started," *Wired*, April 13, 2023, https://www.wired.com/story/chatgpt-jailbreak-generative-ai-hacking/.

[152] Anthropic, https://www.anthropic.com/company. See also "We all need to join in a race for AI safety," Anthropic, July 21, 2023, https://twitter.com/AnthropicAI/status/1682410227373838338.

[153] "Core Views on AI Safety: When Why, What, and How," Anthropic, March 8, 2023, https://www.anthropic.com/index/core-views-on-ai-safety.

[154] Erik Gartzke, Quan Li, and Charles Boehmer, "Investing in the Peace: Economic Interdependence and International Cooperation," *International Organization* 55, Issue 2 (2001): 391-438. https://www.cambridge.org/core/journals/international-organization/article/abs/investing-in-the-peace-economic-interdependence-and-international-conflict/2858A70A728D88ADAE184768537B41E6.

[155] Jervis, *The Logic of Images in International Relations*, 123-132.

[156] Anthony Corso et al., "A Holistic Assessment of the Reliability of Machine Learning Systems," *arXiv*, July 29, 2023, https://arxiv.org/abs/2307.10586.

[157] Steven Feldstein, *The Rise of Digital Repression: How Technology is Reshaping Power, Politics, and Resistance* (Oxford: Oxford University Press, 2021), https://academic.oup.com/book/39418.

[158] The authors are grateful to Jason Brown for his insights on this topic.

[159] Jack Corrigan, Melissa Flagg, Dewey Murdick, "The Policy Playbook: Building a Systems-Oriented Approach to Technology and National Security Policy," Center for Security and Emerging Technology, June 2023, https://cset.georgetown.edu/publication/the-policy-playbook/. For example, Canada's "Responsible Use of Artificial Intelligence" document highlights this tension in its guiding principles: "be as open as we can by sharing source code, training data, and other relevant information, all while protecting personal information, system integration, and national security and defence." See: "Responsible Use of Artificial Intelligence," Government of Canada, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html.

[160] "How to Audit an AI Model (Part I)," OpenMined, July 1, 2023, https://blog.openmined.org/ai-audit-part-1/#:~:text=To%20do%20this%2C%20an%20owner,for%20use%20in%20AI%20audits.

[161] "How to Audit an AI Model (Part I)," OpenMined; Andrew Trask et al. "Beyond Privacy Trade-offs with Structured Transparency," *arXiv*, December 15, 2020, https://arxiv.org/abs/2012.08347; Imbrie et al., "Privacy is Power."

[162] The authors are grateful to Dr. Erik Lin-Greenberg for his insights on this point and for raising helpful questions around the variables that may influence the signaling logics of different technologies.

[163] Micah Musser et al., "Adversarial Machine Learning and Cybersecurity: Risks, Challenges, and Implications," Center for Security and Emerging Technology, April 2023, https://cset.georgetown.edu/publication/adversarial-machine-learning-and-cybersecurity/.

[164] "Failure Modes in Machine Learning," Microsoft, November 2, 2022, https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning.

[165] Heather Roff, "Deception: When Your Artificial Intelligence Learns to Lie," IEEE Spectrum, February 24, 2020, https://spectrum.ieee.org/ai-deception-when-your-ai-learns-to-lie; Matthew Burtell and

Thomas Woodside, "Artificial Influence: An Analysis of AI-Driven Persuasion," *arXiv*, March 15, 2023, https://arxiv.org/abs/2303.08721.

[166] Musser et al., "Adversarial Machine Learning and Cybersecurity."

[167] See, for example, Fiona S. Cunningham, "Cooperation under Asymmetry? The Future of US-China Nuclear Relations," *The Washington Quarterly* 44, Issue 2 (2021): 159-180, https://www.tandfonline.com/doi/abs/10.1080/0163660X.2021.1934253; Ezra Klein Interviews Fareed Zakaria, The Ezra Klein Show, *New York Times*, March 4, 2022, https://www.nytimes.com/2022/03/04/podcasts/transcript-ezra-klein-interviews-fareed-zakaria.html.