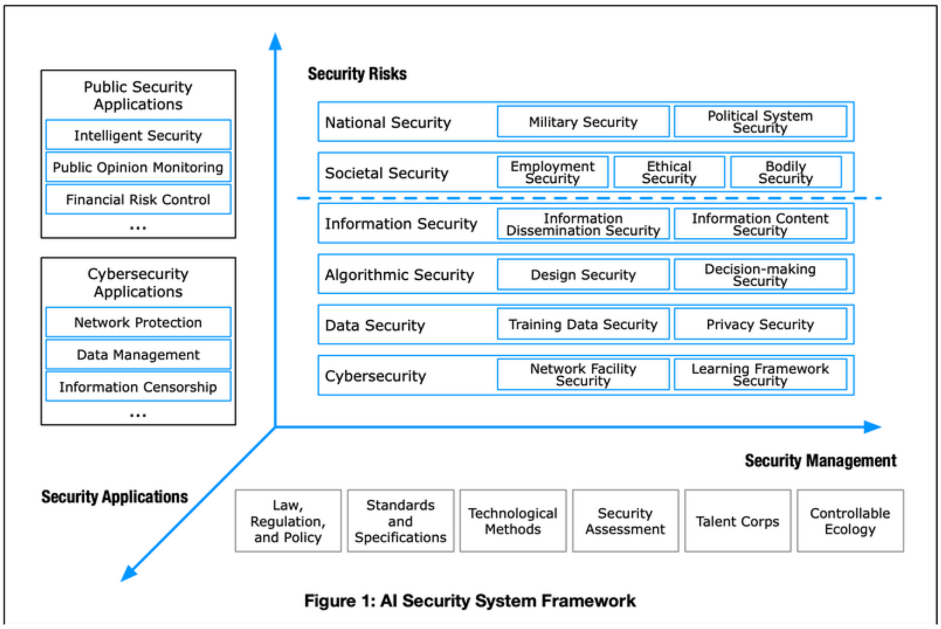
Online Symposium: Chinese Thinking on AI Security in Comparative Context

BLOG POST



CAICT / Translated by DigiChina

By **Paul Triolo**, **Elsa Kania**, **Jacqueline Musiitwa**, **Maarten Van Horenbeeck**, **Justin Sherman**, and **Graham Webster**

Feb. 21, 2019

DigiChina convened its contributor community, the broader New America Cybersecurity Initiative network, and our colleagues at other institutions to assess an important Chinese document on artificial intelligence (AI) policy. We asked commentators to respond to the prompt below, and provided them with the first section of the translated AI Security White Paper, from which further material is provided in **our now-published translation**.

Prompt:

A major government-affiliated think tank in China last year released a wide-ranging white paper on artificial intelligence and security. The paper, from the China Academy of Information and Communications Technology (CAICT), which DigiChina has **previously profiled**, offers a multidimensional assessment of security and safety risks and potential ways to respond to them related to AI.

The authors propose a framework in specific reference to international discussions on AI ethics and governance, ranging from Asimov's laws of robotics to the Asimolar Principles. What does this proposal from Chinese researchers have to offer to the global conversation on AI, and how should the international community respond?

DigiChina, a project of New America, and its partner the Ethics and Governance of AI Initiative of the Berkman Klein Center for Internet and Society at Harvard University and the Massachusetts Institute of Technology Media Lab, have put this question to a community of scholars and practitioners working on technology and society. Some specialize in the ethics and governance of AI with a limited background on China; some specialize in science and technology in China with a limited background in AI.

A note on translation: The paper uses the term 人工智能安全 (réngōng zhìnéng ānquán) throughout, including in its title. "AI safety" and "AI security" would both be appropriate translations in isolation. In the context of this paper, we have used "AI security" to reflect the broad scope of the subject matter discussed, ranging from national security to societal stability, as well as those challenges often discussed under the banner "AI safety" in the United States.

Responses

Paul Triolo

Geo-Technology Practice Head, Eurasia Group, and China Digital Economy Fellow, New America

&
Charlotte Stix

Research Associate and Policy Officer, Leverhulme Centre for the Future of Intelligence, University of Cambridge

The CAICT document on AI and security/safety represents a broad and well thought out effort by a key Chinese government technology think tank to assess the key issues and propose some steps for going forward. It is important to take a step back and look at the broader context within which this document was released. Since China released the national **New Generation AI Development Plan** (AIDP) in July 2017, issues around safety/security and ethics have begun to gain traction within China's research establishment around AI, and within the broader private sector, which is very much driving China's AI sector. Clearly, as called for in the AIDP, Beijing is putting together the pieces of a regulatory system for AI governance, including hot button issues such as safety, security, and ethics.

As part of this process, in early 2018 China's standards organizations began to seriously address governance issues around AI, with the release of a white paper whose drafting was overseen by the China Electronics Standards Institute (CESI). (See the paper and analysis **here**). That paper was a collaborative effort among a number of government think tanks involved with AI research and policy, and leading private sector companies, both big players and smaller companies part of a dynamic group of AI startups, including iFlytek, Huawei, Alibaba, Tencent, and Sensetime.

In addition, in late 2017 and early 2018, the Chinese government announced the formation of two key advisory groups: the New Generation AI Strategic Advisory Committee (for full list see **here**), and a China AI standardization general group and consulting team. The strategic advisory team is dominating by senior academicians, with some input from Chinese private sector companies. The standards general group includes a much broader array of companies, while the consulting team is derived primarily from academic, government think talk, and even defense industry representatives.

In sum, the CAICT paper comes amid a growing competition among different government organizations and advisory groups seeking to establish a primary role in the emerging field of AI governance. Meanwhile, there is no regulatory body in China that clearly "owns" AI technology oversight—a task that cuts across virtually all industrial sectors in China and other areas such as transportation, medicine, and e-commerce. As an **Ministry of Industry and Information Technology element charged with overseeing standards and applications in the telecommunications sector**, CAICT has responsibilities around AI applications primarily in the ICT sector. The CESI effort was aimed more broadly, but that organization has traditionally handled technical standards, not the ethical and safety issues raised in the AI context in this white paper.

In western countries, a similar dynamic is at work as governments grapple with how best to address issues around AI safety, security, and ethics—and how to integrate private sector and NGO knowledge. In western countries, unlike in China of course, there has been substantial sensitivity to the role of governments in this process.

The Partnership on AI, for example, includes a large cross-section of both leading AI technology companies and NGOs, has a Safety-Critical AI working group that is addressing many of the same issues raised in both the CESI and CAICT white papers. The PAI effort, however, does not include any government organizations. When it comes to the consideration of principles and values for AI, the European Union is relatively well placed with its High-level Experts Group on AI (AI HLEG). This group is independent of the European Commission (EC) but acts as an advisory group on policy and investment as well as on the ethical implications of AI. To that end, the AI HLEG is currently working on "Ethics Guidelines for Trustworthy AI," which are to be published in the coming weeks. A **draft version of this document** has been available for public feedback since December 2018 and has received over 500 submissions. The AI HLEG, made up of 52 subject experts, represents a cross-section of relevant stakeholder, including members of private sector companies, academia and NGOs. Furthermore, the European Union is trying to harness public opinion and bring a more diverse set of voices to this discussion via the AI Alliance, a platform where all of society can provide feedback on the progress of the AI HLEG and raise potential concerns.

One of the major differences in the Chinese context is the lack of true NGOs to participate in broader discussions around these issues, and the realization in government sectoral standards and technical organizations that AI applications already or will increasingly play a role in their sectors. Hence in China the government will be involved by definition. This will complicate efforts to gain broader international consensus around some key safety, security, and ethics issues. With the era of autonomous vehicles just around the corner, this will be an increasingly pressing problem to resolve.

Elsa B. Kania

Adjunct Senior Fellow, Center for a New American Security

Beyond their enthusiasm about the positive potential of AI, Chinese technical leaders and policymakers are also starting to engage with concerns over the safety and security implications of rapid advances in AI technologies, which are recognized as a "double-edged sword." As this white paper describes, rigorous and sophisticated consideration of a range of risks and issues that might arise with advances in AI is underway at CAICT, which has emerged as a key player on these issues. Certain aspects of this white paper should hardly be surprising to those who have tracked recent debates on AI safety, including concerns over data security and cybersecurity, as well as defects in algorithms. However, I'd highlight in particular certain elements of this discussion and framework that reveal the extent to which there can be an ideological dimension to the Chinese government's approach to these issues, raising concerns about the impact of China's aspirations for leadership in AI for the future of these technologies. Hopefully, there will be opportunities for the U.S. and Chinese research communities, and even governments, to engage and collaborate on issues of AI safety and security, but potential asymmetries in priorities and concepts are a challenge that should be openly acknowledged from the start.

In particular, the white paper identified risks to state/national security from AI that include not only military concerns but also the security of China's political system, including "hidden dangers" of the impact of AI on public opinion. There is also a focus on the utility of AI to facilitate information censorship and public opinion monitoring. For instance, the paper says, AI can facilitate rapid review and removal of "harmful network content," a notion that is fairly expansively interpreted in the Chinese system. At the same time, AI will be leveraged as the Chinese government looks to improve its capacity for "social governance," which in practice often includes coercive attempts to shape and control society. So too, notions of "societal security" are difficult to extricate from the political and ideological environment in which these concepts are applied. Even concerns about structural employment resulting from AI applications, which are common across countries, will likely be approached by the Chinese government as threats to political/regime security.

Going forward, it will be interesting to see whether the concerns over the risks posed by AI to national security will extend beyond such technical discussions to shape the Chinese military's approach to its own research and development of AI applications. This white paper raises concerns that AI "can be used to build new-type military strike forces, directly threatening national security," and it points out: "the applications of intelligent weapons will cause: control to become remote, increased precision of strikes, miniaturization of conflict domains, and process intelligentization." However, the discussion of trends toward "a new round of arms race" highlights U.S. and Russian efforts without acknowledging the PLA's own extensive investments and developments in the advancement of military intelligentization, which Xi Jinping personally has urged the PLA to advance. As concerns of AI ethics and safety emerge as a core element of the U.S. Department of Defense's own AI Strategy, perhaps China will consider providing greater transparency on the extent to which these concerns influence its own approach to AI for national defense, beyond this initial consideration of such issues by CAICT.

Rui Zhong

Program Assistant, Kissinger Institute on China and the United States, Woodrow Wilson International Center for Scholars

This document establishes frameworks not only to develop AI, but also to manage the teams working on it and the users impacted by emerging technologies. Within the white paper, a definition of national security that includes both military and political system security was identified as a core priority of AI maturation. This definition provides an important indicator of these technologies' potential applications. What regulators seek to curb in AI system developments are any factors that destabilize, interrupt, or impede the operation of government and Chinese Communist Party functions.

Chinese authorities' intentions to use computing technology in policing, combined with Xi Jinping's stated intention to use technology in cleaning up discourse, puts ideological work squarely in the path of AI development trajectories. As solutions that can more efficiently process user data are developed, they will likely go through filters and checks by Party-side officials, including cells installed within China's biggest AI incubators. These teams will likely be working to ensure compliance in security applications, specifically "information censorship applications" and "public opinion monitoring applications" in communications-related AI tools. Both are tasks that carry on existing policy priorities in traditional media oversight and censorship of the Chinese general public's online content and commentary.

The international community should expect the release of additional draft policies that detail specific content guidelines for the assessment and detection of security risks that AI should be able to detect, particularly in public-facing AI tools. These forthcoming details will be key indicators of regulator and firm intentions for Chinese-operated media and communications spaces domestically and abroad.

Jessica Cussins Newman

Research Fellow, Center for Long-Term Cybersecurity, University of California at Berkeley

The AI Security System Framework from CAICT appears to be a useful tool to conceptualize a fairly broad range of AI security implications that goes beyond the usual suspects of national security and cybersecurity. For example, the inclusion of societal security risks could help expose systemic social threats from AI that do not always receive sufficient attention from national authorities, but which could be massively destabilizing to nations and regions of the world.

Jessica Cussins Newman recently published a paper highly relevant to the broader discussion of comparative AI security policy: ***Toward AI Security: Global Aspirations for a More Resilient Future*** *–Ed.*

Jacqueline Musiitwa

Cybersecurity Policy Fellow, New America

Amidst the battle for AI advancement between the United States and China, there are other developments in Africa and priorities globally that are likely to attract more Chinese support.

For instance, in Africa, the discourse is steered less toward AI security issues and more toward the value of AI for development. First, there is a need to shift the stages for research excellence to also include African cities and universities. Google's AI lab in Ghana demonstrates that an African location is worth investing in. Second, STEM education should be promoted. Children need to be raised in an educational environment that fosters curiosity. There is also an opportunity for AI solutions providers to engage in skill transfers.

Third, legal systems across the continent are updating their laws to accommodate such changes.

The challenges of access to AI facilities will continue until government and the private sector collaborate to find solutions. China has been a strong partner to African countries in several areas, including technology transfer and so support in AI is a natural progression.

Maarten Van Horenbeeck

Cybersecurity Policy Fellow, New America

What makes this framework interesting is less what is included than what is missing. The framework published by CAICT focuses on those challenges that can be addressed while increasing use of AI, while mostly avoiding the issues of privacy and individual rights, that more likely to restrict growth.

Speaking to attendees of the Fortune Global Tech Forum in Guangzhou last November, AI expert Kai-Fu Lee noted how AI developments in the United States have been focused on the identification of algorithmic breakthroughs, whereas in China they focus on putting the technology to use by leveraging larger datasets.

This framework aligns with that view by showing a strong focus on the identification and remediation of technical risks posed by AI, as well as ensuring the outcomes of AI can be understood, managed, and controlled—but it doesn't go near questions of when and which AI use cases are appropriate from an individual rights perspective. Privacy is mentioned in the context of disclosure of data, rather than the risks of mass collection and abuse.

In that sense it widens the gap with the European perspective, where privacy is a key element highlighted in draft AI Ethics guidelines, to be published in final form by an EU High Level Expert Group in March.

In the United States, the outcomes of the White House Summit on AI in May 2018 only mentioned privacy once—and only in the context of government data. Under the Networking and Information Technology Research and Development (NITRD) program, ethics and privacy were called out as a priority in 2016. Neither that initiative nor the summit document have resulted in a coherent high-level strategy. This may change with the new executive order dated February 12, which makes multiple mentions of American values as an important part of growing AI leadership. It, however, makes little available in terms of actual funding or detailed commitments, so the effects are yet to be seen.

Beyond not addressing these concepts, the Chinese strategy paper erodes the idea of privacy by identifying specific use cases for AI in the area of "social governance," public opinion monitoring, and information censorship.

Justin Sherman

Cybersecurity Policy Fellow, New America

Rhetoric continues to circulate in democratic countries that China as a whole has an enormous edge over the United States or the European Union as a whole in AI development due to Chinese society's alleged lack of concern for technology ethics. This potentially dangerous rhetoric (presumably based in large part on the view that China's government engages in pervasive surveillance) is often used to conclude that efforts to build safe or ethical AI are futile, and that democratic countries only hinder their AI development by paying attention to algorithmic vulnerability or algorithmic bias along the way. This framework shows that saying “China” has absolutely no regard for AI ethics is inaccurate and overly sweeping.

This proposal also demonstrates the intention of Chinese researchers and other entities to engage in global dialogues on AI ethics. And contrary to the zero-sum conclusions that result from the framing of AI competition as an “AI arms race,” there may in fact be many areas of mutual interest between the United States and China regarding safe and ethical AI. Many issues discussed in the paper, such as algorithmic model defects and training data bias, are the same issues raised in democratic discourse around safe and ethical AI. Advancing global norms and standards in these areas may therefore stand to benefit all.

At the same time, this does not mean that certain norms or standards won’t be pushed by the Chinese government, in an attempt to gain an edge in AI development or to further promote digital authoritarianism around the world. The issue is similar to internet governance in this way: While some norms or standards developed in international forums undoubtedly benefit all (e.g. interoperability rules that allow basic internet connectivity), other norms or standards are used as top cover for censorship, surveillance, and tight control of the net (e.g. China’s cyber code of conduct proposals at the U.N. General Assembly). Other norms or standards have disproportionate benefits for certain countries. Democratic policymakers should therefore pay close attention to China’s attempts to set standards or develop norms around the design, construction, training, testing, and deployment of artificial intelligence. The United States is behind China as it is, given that its government only just issued an **executive order** on national AI development. Because AI will have great impacts on state power and the future world order, American policymakers in particular must engage with American researchers on issues of safe and ethical AI while doing the same on the international stage.