The Future of Multilateralism and Global Governance

G20 2020

Coordinating Committee for the Governance of Artificial Intelligence

Thorsten Jelinek (Taihe Institute),
Wendell Wallach (Yale University Interdisciplinary Center for Bioethics),
Danil Kerimi (World Economic Forum (WEF))
June 25, 2020 | Last updated: July 7, 2020

This Policy Brief is offered to the Saudi T20 process, as a recommendation to the G20 in 2020.

This policy brief proposes to the G20 the implementation of a Coordinating Committee for the Governance of Artificial Intelligence (CCGAI) to effectively coordinate on a global level the prevention and mitigation of direct cyber-physical threats and long-term structural risks. The G20 is the appropriate institution for a CCGAI given its influence on international policy. The CCGAI requires further institutionalization of the G20 to increase trust and legitimize such global umbrella role, while effectively countering today’s digital regime complex fragmentation. The challenges related to AI governance, the institutional features of the CCGAI, and an initial agenda including the most urgent topics are highlighted in this policy brief, including the proposal of an informal, coordinating forum that could function as a forerunner and light version of the CCGAI.

Challenge

There is an urgent need for global coordination of the governance of AI [26]. Automated decision-making, coupled with the reuse of mass data and ubiquitous digitalization, has become a global driver for economic and strategic competitiveness. However, no single country or stakeholder can effectively and sustainably prevent and mitigate the changing landscape of direct cyber-physical threats and longer-term structural risks that will impact entire societies, economies, and governments, as well as international and strategic relations [2]. AI applications span a broad array of domains and pose unique security threats in each. Such technology determinism ultimately raises fundamental questions concerning human dignity and existence and therefore need to be addressed on a global level [28].

The proliferation of normative frameworks, which advocate for a responsible or human-centered use of AI, reveal the widespread perception of a fundamental normative and governance gap [15, 27]. While those frameworks have been defined fairly rapidly, respective governance approaches, which designate possibilities for collaboration and should be guided by those normative commitments, are still lacking and will be much more difficult to realize. There are at least three fundamental dynamics that undermine the governance of AI and, conversely, make a global coordinating mechanism an urgent necessity [13]:

First, AI is based on disparate technologies that have different threat and risk scenarios across different applications, sectors, and geographies. Those technologies are advancing and being deployed rapidly, will eventual permeate all aspects of human life. Existing regulations and traditional regulatory approaches do not match such complexity, nor can they keep up with the speed of AI’s advancement
and adaptation [26]. Second, AI governance, which includes coordinated actions concerning ethics, norms, policies, industry standards, laboratory practices, and engineering solutions, is exposed to fierce competition over global AI leadership. Competition fosters innovation, but also compromises responsibility and leads to a concentration of AI resources and power imbalances. Third, cultural differences and competing political interests and governmental systems, especially today’s big power competition, lead to conflicting normative frameworks and regulations. They increase tension between state actors and further undermine much-needed international cooperation. Those differences and tensions are perpetuated through rising nationalism and populism and a heightened distrust in multilateralism [13, 19].

AI increasingly amplifies the broader discourse on digitalization and cyberspace, which already manifests as a highly fragmented “regime complex” [20]. Without global coordination and joint interventions, the increasing demand for “digital sovereignty” could turn into “technological nationalism” and reinforce a low-trust environment. AI bears its own technological risks, but it is human behavior and the use of AI that primarily risk reinforcing the current trajectory of humankind, as and history has entered the downward spiral of “contested multilateralism” and “great power competition,” the risk of experiencing more of the downside of AI and technology determinism is likely [13, 19, 28]. A globally disruptive trend within an already fragmented environment requires a globally coordinated response. The Group of Twenty (G20) is the obvious institution to implement a Coordinating Committee for the Governance of Artificial Intelligence (CCGAI) due to the group’s considerable influence on international policy coordination and framework design [12].

Proposal

Balancing the need for innovation, competition, and cooperation while mitigating the risks and undesirable consequences attributed to AI poses a daunting challenge for governments. This challenge arises from the dual-use, uncertain, and increasingly all-embracing character of AI-driven digitalization and robotics, as well as by an already fragmented cyber regime complex and the increasing lack of international cooperation and trust [19, 20]. Therefore, this policy brief urgently proposes the implementation of an G20 CCGAI [cf. 26]. In 2019, the G20 agreed on a set of norms for “human-centered AI that promotes innovation and investment” [11]. The G20 should build on and expand those recommendations, which were derived from the OECD Principles on AI [21], and implement the proposed coordinating mechanism. For the G20, this would be an opportunity to actively reduce and mitigate AI threats and risks, while countering today’s fragmentation through integration, coherence, and respecting differences.

Demand for an International Coordinating Mechanism

The informal organization of a deliberative, international forum by a nonpermanent, rotating secretariat that facilitates loose linkages and groupings between the most powerful state and non-state actors is frequently seen as what has guaranteed the continuation of the G20 since inception. At the same time, such informality and flexibility has also been scrutinized as the G20’s weakness and limitation [1, 25]. The establishment of a G20 CCGAI would demand further institutionalization of the G20 process, but only concerning the issue of AI governance [cf. 4]. In this policy brief, such centralization is deemed necessary to improve the effectiveness not only of the G20 but also of the entire cyber regime complex in reducing and mitigating AI cyber-physical threats and longer-term structural imbalances. Today, the G20 is only one among various actors within a deeply divided cyber regime complex. However, the G20 has the capacity for such global stewardship and a CCGAI could tremendously improve the overall functionality of today’s cyber regime complex.

A proliferation of non- or partially integrated organizational, national, and regional normative and regulatory approaches has been the initial response to this globally emerging technology. There are clear advantages of decentralized, self-organized, polycentric, or network-driven governance arrangements [4, 24]. They are efficient in identifying the wide range of uncertainties, policy issues, and innovative solutions adjusted to local requirements. On a regional level, for example, the European Commission has managed to integrate national responses of member states and developed pre-regulatory AI and data strategies building on and referencing existing EU normative frameworks and laws [6, 7]. Such regional integration is a response to mitigating AI and data risk as well as to enhancing competitiveness. However, on a global scale, the strategic and competitive nature of the cyber space and AI-driven digitalization has largely reinforced a “return to the nation state” [20, p.3]. The demand for digital sovereignty, which seeks balance between protection and collaboration, risks both undermining multilateralism and leading to “digital nationalism” and “technology determination” [14, 19, 28]. The result is a dysfunctional international regime complex that will weaken local and regional approaches and render them ineffective [cf. 16, 20]. Thus, only a comprehensive approach coordinated on a global level is effective to prepare against, mitigate, and recover from future threats and structural imbalances,
and eventually address still distant scenarios of a transhumanist era [26].

A CCGAI does not mean to be a single legal structure with direct enforcement authority and fully integrated international cyber regime complex. Neither would such level of centralization be feasible nor desirable given the nature and advantages in the formation of bottom-up and self-organized regimes or governance networks. However, the CCGAI must strive to counter fragmentation by striking a balance between the G20 as an informal and crisis response-driven institution and a G20 that takes on a formal global umbrella role for ongoing cooperation and coordination. Such an umbrella role would build upon and align with established procedures, shared long-term orientations and action plans, and joint presentations and appearances [cf. 12]. The implementation of a CCGAI would require further institutionalization of the G20 based on, but not limited to, the following four institutional features that would mandate the CCGAI as a "metagovernor": coordination, accountability, foresight, and consultation [cf. 1, 4, 12, 22, 23]:

1. **Comprehensive coordination** is a "metagovernance" [14] task to build and institutionalize linkages between the CCGAI and relevant actors within the G20 complex, including committees, boards, task forces, and engagement groups such as the B20, C20, and T20. The overall task is to synchronize, integrate, and delegate responsibilities and decision-making between the competencies. Equally important, such an empowering coordinating function must also formally build and maintain linkages between the G20 and the main actors and hierarchies within the broader AI and cyber regime complex. In this process, the CCGAI does not seek to compete against other institutions and regimes but to facilitate collaboration with the aim of achieving integration and coherence, and for supporting the implementation of a global agenda for responsible AI governance. The coordination function could serve to prepare and negotiate international agreements and treaties and to help the G20 develop from a discrete into an active agent.

2. **Accountable procedures** are paramount for gaining legitimacy and trust [3, 22, 23, 24]. Coordinating between member states, competencies, hierarchies, and governance networks and reaching decisions require transparent, rules-based, justifiable, and sanctionable procedures. Such formalization is crucial, but it is not transparency alone that contributes to the effectiveness of the CCGAI. Coordination must also remain flexible and leave space for informality, both of which have contributed to the continuation of the G20. As consensus will not always be feasible within the current fragmented context and with uncertain technology, the CCGAI must also follow a normative procedure for tolerating and managing ambiguity and conflict. The CCGAI should look for common views, respect differences, and facilitate debate over differences in hopes of forging common views over time [4].

3. **Strategic foresight** allows for improving the effectiveness of coordination and decision-making [4]. It requires monitoring the development and application of AI and related policies, incubating and accelerating policy responses, and proposing early warnings and international mitigation strategies in relation to a continuously updated spectrum of AI threats and risks. In this process, the CCGAI would not primarily promulgate new governance instruments; rather, it would share oversight outcomes and catalyze the instruments that have already been promulgated or proposed. The CCGAI could analyze how existing governance and regulatory instruments fit together, where they agree, and where gaps and policy conflicts are left that need to be addressed. Strategic foresight should also be a function to measure the CCGAI's own capability to lead and improve the functionality of the AI and cyber regime complex based on the following six criteria: coherence, accountability, effectiveness, determinacy, sustainability, and epistemic quality [15]. Foresight information should be stored in the already existing G20 Repository of Digital Policies [7, p.2, 10].

4. **Public consultation** improves the transparency and effectiveness of the governance coordination process and creates legitimacy and trust [1, 3, 4]. A consultation mechanism needs to be formalized where stakeholders, especially civil society groups and non-G20 countries, are integrated into a separate secretariat and contribute at the level of official policy discussions. Public consultation is a platform for providing feedback, raising concerns, and addressing asymmetric power relations and domination, including the needs of small nations and underserved communities. It should be an instrument that enables an inclusive coordination process, empowers self-organization and governance networks, and helps to accommodate a multilayered, multidisciplinary, and polycentric environment. Fair access instead of preferential treatment must be provided. Public consultation is a mechanism for true multistakeholder input and allows the CCGAI to remain open, flexible, and adaptive. The CCGAI should collaborate with other...
multistakeholder input, and allows the G20 to remain open, flexible, and reflexive. The CCGAI should collaborate with other organizations and, in particular, the United Nations, which has already established strong links with civil society and runs programs for digital cooperation and governance.

**Coordinating Forum as Intermediate Step**

An ad-hoc implementation of those institutional features requires a consensus among the G20 member states as well as the acquisition of additional resources to plan, implement, and operate a CCGAI. Asking for such commitment might therefore prevent the establishment of a CCGAI. Hence, this policy brief also proposes a Coordinating Forum for the Governance of AI (CFGAI) that could function as an intermediate step towards the establishment of a CCGAI. Such a light version of a CCGAI would not require any reform of the G20, but would invite major stakeholders to discuss the goals, principles, and institutions of a future coordinating committee as well as the threat and risk scenarios and themes outlined within this policy brief. Participation would be on a voluntary, but recurring basis to ensure a seamless continuation of the debate and follow up with joint declarations and tasks. The CFGAI should be understood as a precursor, subsequently testing and implementing new institutions and, ultimately, leading to the establishment of a coordinating committee.

**Prevention and Mitigation of Direct Threats and Structural Imbalances**

The above section outlines the normative aspects of international coordination or metagovernance. For coordination to operate clearly and effectively, it is necessary to specify the object of coordination itself, which is the different sectors, dimensions and specific aspects of AI norms, governance, and engineering. The joint target of coordination and policy discussions involves at least a common definition of the AI technology [cf. 5], the broader AI ecosystem [cf. 17], and the risk profile [cf. 2]. There are various definitions of each of those domains, which need to be revisited, and a common understanding needs to be reached and frequently updated by the CCGAI. This policy brief draws the focus to the latter, a comprehensive AI risk profile [2, 13], which should be at the center of prioritizing and structuring international coordination and support realizing the G20’s commitment to human-centered AI. The use of AI has been cautioned against as a source of unprecedented risks. Those risks can be clustered into two groups [13, 2]: (a) threats that are experienced directly in a specific domain and (b) risks that are structural and unfold over a longer period of time.

1. **Direct threats**: The advancement and diffusion of AI technologies impact the existing landscape of cybersecurity threats. Cyber threats will change and intensify tremendously due to the adversarial use of AI. There will be an expansion of existing threats, more effective and targeted threats, and the emergence of entirely new types of cyber-physical threats. In addition to such intended attacks—causing disruption, theft, or espionage—there will be unintended and unpredictable accidents, which will also be the target of intentional exploitation. Against such an intensifying scenario of cyber-physical threats, the question of AI security has already become a matter of national security and protecting critical national infrastructure. Without stronger commitment for international coordination and responsibility, AI security questions might further divide and fragment the cyber regime complex.

2. **Structural imbalances**: Without coordination and intervention, AI-driven digitalization risks cause severe structural imbalances. Structural risks have longer-term consequences, which are more difficult to anticipate and mitigate, but their impact is expected to be much more widespread and pervasive. As technology is an integral part of and not external to human behavior, the use of AI strongly risks reinforcing current historical developments. The structural risk will impact all fundamental dimensions of human affairs, including the economy, society, politics, international relations, and geopolitical security. Economically, mass labor displacement, underemployment, and de-skilling are likely outcomes, which especially threaten low- and middle-income countries. For societies, increasing lack of dignity, privacy, and meaning will threaten both physical and psychological well-being and social cohesion. Politically, AI increases the structural risk of shifting the power balance between the state, the economy, and society by limiting the space for autonomy. While authoritarian states could slide into totalitarian regimes, democracies could witness the erosion of their institutions and the disintegration of public morality and manufacturing consent from the governed. A fierce global competition over AI leadership risks disrupting existing international relations. Technology sovereignty could turn into technology nationalism and enable political capitalism. Ultimately, proliferation and easy accessibility of offensive, AI-enabled cyber capabilities, notably lethal autonomous weapons, increases the risk of ongoing asymmetric conflicts.
The CCGAI needs to monitor and map the full spectrum of direct threats and structural risks and understand the emerging interdependencies between the use of AI technologies and the broader dimensions of human affairs. Although security is generally not a domain of the G20, AI security should be included given its risk of reinforcing the division and fragmentation of the cyber regime complex.

The purpose of such comprehensive monitoring is not only to inform and direct policy discussions but also to coordinate and develop international mitigation strategies, early warning systems, and crisis response plans. Derived from the risk spectrum outlined above, the following themes for global coordination are proposed that will be impacted by AI.

1. **Digital sovereignty**: policies balancing between digital and technology sovereignty and multilateralism and a global level playing field.

2. **Inclusive digital economy**: ensuring a just transformation of work and society, while promoting AI and data as drivers for a digital global economy, innovation, and competitiveness.

3. **Market power imbalances**: addressing the needs of developing nations and underserved communities through capacity building and adaptation of development models.

4. **International security**: possible conventions, roles and responsibilities in cyberspace concerning the proliferation of offensive cyber technologies.

5. **System failures**: minimizing and mitigating the risks of unintended system failures and exploitations of engineering loopholes.

6. **AI for common good**: utilizing technology for the common good, including areas such as decarbonization, health and pandemics, energy, food, and inequality.

7. **Coordination architecture**: as governance failure is a primary risk itself, coordination and governance mechanisms must remain part of ongoing discussions and reform.

### Organization and Cooperation

The G20 AI coordinating mechanism should comprise a coordinating committee, advisory group, working group, cooperation accelerator, policy incubator, and observatory with foresight and help desk capacity. As the highest-level body, the coordinating committee should be a permanent, chartered committee, led by annually rotating co-chairs, and convene the heads of state and government and key non-state representatives. Its members must agree on common objectives and norms, designing and implementing the coordinating mechanism, defining and adhering to the criteria for the functionality of the CCGAI, including coherence, accountability, effectiveness, determinacy, sustainability, and epistemic quality. The coordinating committee encourages and follows the institutional features of the CCGAI, like those four features proposed above. The members of the committee seek consensus, make recommendations, and agree upon coordinated plans and actions but need to remain respectful of differences. The CCGAI must seek to continuously improve itself as an agile, cooperative, and comprehensive international coordinating mechanism [cf. 26]. Initially, the CCGAI should agree upon a common charter that captures the joint commitment of the member states as well as the overall goals and procedures of the CCGAI.

The G20 needs to build its own coordination and implementation capacity to carry out the function of a CCGAI, should incorporate related work that has been done within the G20 complex, and establish linkages to existing procedures, declarations, principles, and tools. Notably, it should revisit the Digital Economy Development and Cooperation Initiative (China 2016) [8], Digital Economy Ministerial Declaration (Germany 2017) [9], and Ministerial Statement on Trade and Digital Economy and AI Principles (Japan 2019) [11] and utilize the G20 Repository of Digital Policies (Argentina 2018) [10]. However, the CCGAI cannot and must not own and carry out all of the proposed functions and topics. Some of them should be carried out by external organizations and regimes, but the CCGAI should remain the primary coordinating body. Regarding the meta-governance function and selected thematic areas, the CCGAI should closely collaborate with the United Nations and other multilateral institutions.

### Obstacles to the Coordinating Committee

Regimes are usually initiated and maintained by the most powerful states. Yet, big power competition and increasing nationalism and populism, as well as the disruption of the post-war liberal order, likely undermine establishment of such an international coordinating mechanism due to the fear of compromising influence and power [4]. In addition, the low-trust environment most likely remains an enduring condition, and fierce competition and political and cultural conflict reinforce self-interest and fragmentation [20]. Additionally, the private sector might resist as large businesses struggle to maintain their privileged and informal access to the G20 [18]. Furthermore, there has been an ongoing resistance within the G20 to reforming itself and turning into an accountable, rules-based, and treaty-bound organization with a
an ongoing resistance within the G20 to reforming itself and turning into an accountable, rules-based, and treaty-bound organization with a permanent secretariat. However, the G20 was established due to the rise of the multipolar world and middle-power countries, and it is those countries that have a strong interest in multilateralism and the use of the G20 meeting as the forerunner of a CFGAI and CCGAI, respectively. To balance private sector interests and help increasing the trust in businesses and institutions, the G20 should forge public-private partnerships that provide a vision for a "caring economy" [cf. 29] and help to ensure that essential resources are managed as "commons" [30].

References


Existing Initiatives & Analysis