QUT Centre for Data Science

Draft G7 Guiding Principles for Organizations Developing Advanced AI systems

QUT Centre For Data Science submission

LEAD AUTHOR:

Aaron Snoswell

CONTRIBUTING AUTHORS:

Bernadette Hyland-Wood, Kerrie Mengersen, Tamara Orth, Becki Cook, Bemali Wickramanayake

SUBMISSION DATE:

20 October 2023

1. About the Queensland University of Technology Centre for Data Science

Data science is a powerful force for addressing the challenges we face across all sectors — health, environment, business, society and industry. Its ability to solve global challenges is at the forefront of discussions and strategic activity in most commercial and government organisations, research entities and universities.

The Queensland University of Technology (QUT) Centre for Data Science (CDS, https://www.qut.edu.au/research/centre-for-data-science) brings together QUT's strengths and track record in data analysis, computation and eResearch, as well as cross-disciplinary projects and connections to centres of excellence, cooperative research centres, and national and international networks. The CDS also draws together capability in data science from across Australia, providing a centralised hub for world-class data science research, unique training opportunities, and active external engagement.

At the CDS, we leverage our diverse capabilities in data science to deliver world-class research, unique training opportunities and active external engagement through:

1. Development of frontier methods in data science, solving fundamental problems in gathering, modelling and analysing data, to make data-informed decisions that deliver benefits to industry, society and the environment.
2. Connecting the breadth and depth of data science researcher across all of QUT, to develop state-of-the-art solutions, and
3. Connect people, resources and opportunities across the many Australian universities and industry partners working with data through our broader national engagement with the Australian Data Science Network (ASDN, https://www.australiandatascience.net/).

2. Our interest in the G7 Guiding Principles for Organizations Developing Advanced AI

Responsible Data Science forms one of the five core programs in the CDS's research agenda. The Responsible Data Science Program draws on a range of nationally and internationally recognised research expertise in AI, ethical and explainable AI (XAI), data governance, data sovereignty, data federation, digital media research and technology law. Within the context of applied data science, the Program engages on issues of national and global significance, bringing together traditional and new voices on approaches, standards, and best practices in support for trustworthy data.

By convening participants from industry, government, academia, and civil society in focused, time bounded collaborations, the program aims to build Australia's expertise and engagement in responsible data collection and use through active engagement, and a portfolio of research projects, programming, and educational opportunities.

Australia has to-date adopted a proactive role in shaping AI policy nationally and beyond our borders (for instance, through the Australian AI Framework, and AI ethics principles, published by the Department of Industry in 2019). Although Australia is not a G7 or EU member, we share strong international connections with G7 and EU members, and have an interest in promoting responsible research, development, deployment, and distribution of AI systems, wherever they are physically situated. Furthermore, AI systems are not isolated artefacts, but are increasingly embedded in global geo-political and logistical logics, which further emphasises the need for international engagement and collaboration to address the potential risks and opportunities from advanced AI systems.

3. Submission Responses

Question 1. To what type of organizations employing AI should those guiding principles apply?

Developer - an organisation that designs, codes, or produce AI systems); Deployer - an organisation that uses AI systems; Distributor - an organisation other than the developer that makes an AI system available on the market.

[X] Developers
[X] Deployers
[X] Distributors
[X] other actors in the AI value chain (please specify below) – Research institutions.

We emphasise here that often in the AI field the line between research and development is blurred. To avoid organizations 'dodging' these voluntary principles, we recommend that these principles also be applicable to research institutions working on advanced AI systems.

Question 2. Please indicate how important you consider each of the following principles on a scale from 1 (= not very important) to 5 (= very important)

Please evaluate the importance of each principle to achieve the main objective of providing effective and proportionate guardrails on advanced AI systems on the global level and solicit views on any possibly missing elements or actions.

Principle 1: Take appropriate measures throughout the development of advanced AI systems, including prior to and throughout their deployment and placement on the market, to identify, evaluate, mitigate risks across the AI lifecycle.

Very important (5)

Principle 2: Identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment including placement on the market.

Very important (5)

Principle 3: Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency.

Very important (5)

Principle 4: Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia.

Important (4) – But note the need for responsible sharing to avoid exacerbating vulnerabilities. It may be appropriate to reference responsible disclosure practices developed in the cyber security discipline here (see, e.g. [4]).

Principle 5: Develop, implement and disclose AI governance and risk management policies, grounded in a risk-based approach – including privacy policies, and mitigation measures, in particular for organizations developing advanced AI systems.

Important (4) – But noting that while a risk-based approach is concordant with the proposed EU AI Act, risk-based frameworks may not be adopted by all nations or all G7 members. Risk-based regulatory approaches, used in isolation, can have limitations (see e.g. responses to questions 14, 15, and 17 in the ADM+S Centre's response to the Australian Government's Responsible AI discussion paper [5]).

Principle 6: Invest in and implement robust security controls, including physical security, cybersecurity and insider threat safeguards across the AI lifecycle.

Neither important or not (3) – This principle has no specific bearing on AI system development, but is a generally applicable digital security principle. The wording of the principle around 'securing model weights and algorithms' is also concerning as this seems in opposition to open-source development of advanced AI systems. There are benefits to both proprietary and open-source AI development pathways, and voluntary guiding principles such as these should not necessarily encourage one approach at the expense of the other.

Principle 7: Develop and deploy reliable content authentication and provenance mechanisms such as watermarking or other techniques to enable users to identify AI-generated content.

Neither important or not (3) – We agree that the problem of content authenticity and provenance is increasingly challenging and pertinent due to advances in AI systems. However, we note that the reliable identification of AI generated content may not be feasible. Post-hoc detection methods have fundamental accuracy limitations due to universal approximation theorem results [6], and due to 'arms race' dynamics in a free-market environment, and watermarking of AI content is not always appropriate in application contexts, is not always feasible (e.g. in the case of partially AI generated content), and raises surveillance concerns.

Labelling and disclaimers for AI generated content, and so users know when they are interacting with AI systems, are much more generally applicable and feasible near-term ways to partially address the important challenges of content authenticity and provenance.

Principle 8: Prioritize research to mitigate societal, safety and security risks and prioritize investment in effective mitigation measures.

Important (4) – But note that research and development outcomes are not always predictable.

Principle 9: Prioritize the development of advanced AI systems to address the world's greatest challenges, notably but not limited to the climate crisis, global health and education.

Important (4) – But note that research and development outcomes are not always predictable, and some AI systems (such as foundation models) have many down-stream potential applications that may not be immediately obvious. To this end, voluntary guiding principles such as this principle should not discourage fundamental research and development efforts, which may not have immediately obvious applications. We also note that while it is essential to emphasize the development of applications that contribute to societal improvement and sustainability, it is equally crucial to proactively discourage applications that hold the potential to harm society, such as military AI applications.

Principle 10: Advance the development of and, where appropriate, adoption of where appropriate, international technical standards.

Very important (5)

Principle 11: Implement appropriate data input controls and audits.

Very important (5) – But note that 'audits' of input data are mentioned in the principle title, but not elaborated on. This should be amended.

Question 3. In these Guiding Principles, G7 Members also commit to develop proposals to introduce monitoring tools and mechanisms to help organizations stay accountable for the implementation of these actions. How this monitoring mechanism should [sic] look like?

[ ] monitoring by an internationally trusted organisation
[X] national organisations
[X] self-assessment
[ ] no monitoring

International organizations may not be able to take into account the specific contexts of individual nations appropriately. A combination of national organizations (external to AI system researchers/ developers / deployers / distributors) internal self-assessment by AI system researchers / developers / deployers / distributors may be a good approach here.

Question 4. The Draft Guiding Principles include a non-exhaustive list of 11 guiding principles addressed to the AI organisations. It will be discussed and elaborated as a living document that might be completed after Leaders' endorsement, do you have any specific suggestions to add new principles to the list or modify the existing ones?

3000 character(s) maximum

Definitional clarity

Advanced Artificial Intelligence should be carefully defined. E.g., there is a risk that actors might avoid participation in these principles by claiming they are "only doing data science, not AI". The draft guiding principles could be more explicit about what systems and/or applications do or do not constitute non/advanced AI.

AI Dual-Use Nature

The guiding principles acknowledge unacceptable AI uses but overlook AI's dual-use nature. E.g., Generative AI can create 'deepfakes' or foster creative expression. Recognizing this may require risk-balancing [1] rather than categorizing AI systems as strictly acceptable or not.

Safety vs. Societal Risks

The guiding principles disproportionately stress AI system safety, potentially overshadowing more urgent societal risks. E.g. the name "Hiroshima Process" analogizes AI to existential nuclear threats – a fraught comparison [2], and principle 6 overly emphasizes AI safety narratives, potentially inhibiting open-source innovation and distracting from more pressing societal harms [3, 4]. There are benefits to both proprietary and open-source AI development, and guiding principles should not advocate one approach at the expense of the other.

Risk-Based Regulation Limitations

While Principle 5 aligns with the proposed EU AI Act, not all nations or G7 members may adopt risk-based regulatory frameworks. Employing risk-based approaches exclusively can have limitations [6].

Content Authenticity and Provenance

Principle 7 addresses the challenge of content authenticity in AI-generated content. Identifying such content reliably poses difficulties technical and practical difficulties. Using labels and disclaimers for AI-generated content and systems is a more feasible and widely applicable way to address this issue.

Predicting R&D Outcomes

Principles 8 and 9 may not account for the unpredictability of research and development outcomes, especially in AI systems with diverse downstream applications. These principles should not discourage fundamental R&D efforts without immediately obvious applications.

Transparency of System Incentives and Active Education

AI systems have impacts on both individuals and society at large. Disclosure of incentives behind the development of systems can direct foster responsible use and development. Principle 3 addresses public reporting and transparency, but does not address incentives, or account for individual users. Actors should foster active education of users around the systems limitations, risks and possible consequences. In high-stakes domains, AI systems should also (a) have the ability to disclose how decision are made , and (b) act as an enabler of a human decision maker not an autonomous decision agent.

References:

[1] Fraser, Henry, and José-Miguel Bello y Villarino. "Acceptable Risks in Europe's Proposed AI Act: Reasonableness and Other Principles for Deciding How Much Risk Management Is Enough." European Journal of Risk Regulation (2023): 1-16.

[2] Matthews, Dylan. "The AI-nuclear weapons analogy, explained." Vox. (2029).

[3] Spirling, Arthur. "Why open-source generative AI models are an ethical way forward for science." Nature 616.7957 (2023): 413-413.

[4] Castillo, Carlos, et al. "Statement on AI Harms and Policy". ACM FAccT. (2023).

[5] Ķinis, Uldis. "From Responsible Disclosure Policy Towards State-Regulated Responsible Vulnerability Disclosure Procedure: The Latvian Approach." Computer Law & Security Review 34.3 (2018): 508-522.

[6] Weatherall, Kimberlee et al. "Safe and Responsible AI in Australia Discussion Paper: ADM+S Submission" Analysis and Policy Observatory. (2023).