# Executive Summary

In the past several years, seemingly every organization with a connection to technology policy has authored or endorsed a set of principles for AI. As guidelines for ethical, rights-respecting, and socially beneficial AI develop in tandem with – and as rapidly as – the underlying technology, there is an urgent need to understand them, individually and in context. To that end, we analyzed the contents of thirty-six prominent AI principles documents, and in the process, discovered thematic trends that suggest the earliest emergence of sectoral norms.

While each set of principles serves the same basic purpose, to present a vision for the governance of AI, the documents in our dataset are diverse. They vary in their intended audience, composition, scope, and depth. They come from Latin America, East and South Asia, the Middle East, North America, and Europe, and cultural differences doubtless impact their contents. Perhaps most saliently, though, they are authored by different actors: governments and intergovernmental organizations, companies, professional associations, advocacy groups, and multi-stakeholder initiatives. Civil society and multistakeholder documents may serve to set an advocacy agenda or establish a floor for ongoing discussions. National governments' principles are often presented as part of an overall national AI strategy. Many private sector principles appear intended to govern the authoring organization's internal development and use of AI technology, as well as to communicate its goals to other relevant stakeholders including customers and regulators. Given the range of variation across numerous axes, it's all the more surprising that our close study of AI principles documents revealed common themes.

The first substantial aspect of our findings are the eight key themes themselves:

- Privacy. Principles under this theme stand for the idea that AI systems should respect individuals' privacy, both in the use of data for the development of technological systems and by providing impacted people with agency over their data and decisions made with it. Privacy principles are present in 97% of documents in the dataset.
- Accountability. This theme includes principles concerning the importance of mechanisms to ensure that accountability for the impacts of AI systems is appropriately distributed, and that adequate remedies are provided. Accountability principles are present in 97% of documents in the dataset.
- Safety and Security. These principles express requirements that AI systems be safe, performing as intended, and also secure, resistant to being compromised by unauthorized parties. Safety and Security principles are present in 81% of documents in the dataset.
- Transparency and Explainability. Principles under this theme articulate requirements that AI systems be designed and implemented to allow for oversight, including through translation of their operations into intelligible outputs and the provision of information about where, when, and how they are being used. Transparency and Explainability principles are present in 94% of documents in the dataset.
- Fairness and Non-discrimination. With concerns about AI bias already impacting individuals globally, Fairness and Non-discrimination principles call for AI systems to be designed and used to maximize fairness and promote inclusivity. Fairness and Non-discrimination principles are present in 100% of documents in the dataset.

- Human Control of Technology. The principles under this theme require that important decisions remain subject to human review. Human Control of Technology principles are present in 69% of documents in the dataset.
- Professional Responsibility. These principles recognize the vital role that individuals involved in the development and deployment of AI systems play in the systems' impacts, and call on their professionalism and integrity in ensuring that the appropriate stakeholders are consulted and long-term effects are planned for. Professional Responsibility principles are present in 78% of documents in the dataset.
- Promotion of Human Values. Finally, Human Values principles state that the ends to which AI is devoted, and the means by which it is implemented, should correspond with our core values and generally promote humanity's well-being. Promotion of Human Values principles are present in 69% of documents in the dataset.

The second, and perhaps even more striking, side of our findings is that more recent documents tend to cover all eight of these themes, suggesting that the conversation around principled AI is beginning to converge, at least among the communities responsible for the development of these documents. Thus, these themes may represent the "normative core" of a principle-based approach to AI ethics and governance.[1]

However, we caution readers against inferring that, in any individual principles document, broader coverage of the key themes is necessarily better. Context matters. Principles should be understood in their cultural, linguistic, geographic, and organizational context, and some themes will be more relevant to a particular context and audience than others. Moreover, principles are a starting place for governance, not an end. On its own, a set of principles is unlikely to be more than gently persuasive. Its impact is likely to depend on how it is embedded in a larger governance ecosystem, including for instance relevant policies (e.g. AI national plans), laws, regulations, but also professional practices and everyday routines.

One existing governance regime with significant potential relevance to the impacts of AI systems is international human rights law. Scholars, advocates, and professionals have increasingly been attentive to the connection between AI governance and human rights laws and norms,[2] and we observed the impacts of this attention among the principles documents we studied. 64% of our documents contained a reference to human rights, and five documents took international human rights as a framework for their overall effort. Existing mechanisms for the interpretation and protection of human rights may well provide useful input as principles documents are brought to bear on individuals cases and decisions, which will require precise adjudication of standards like "privacy" and "fairness," as well as solutions for complex situations in which separate principles within a single document are in tension with one another.

[1] Both aspects of our findings are observable in the data visualization (p. 8-9) that accompanies this paper.

[2] Hannah Hilligoss, Filippo A. Raso and Vivek Krishnamurthy, 'It's not enough for AI to be "ethical"; it must also be "rights respecting"', Berkman Klein Center Collection (October 2018) https://medium.com/berkman-klein-center/its-not-enough-for-ai-to-be-ethical-it-must-also-be-rights-respecting-b87f7e215b97.

The thirty-six documents in the Principled Artificial Intelligence were curated for variety, with a focus on documents that have been especially visible or influential. As noted above, a range of sectors, geographies, and approaches are represented. Given our subjective sampling method and the fact that the field of ethical and rights-respecting AI is still very much emergent, we expect that perspectives will continue to evolve beyond those reflected here. We hope that this paper and the data visualization that accompanies it can be a resource to advance the conversation on ethical and rightsrespecting AI.