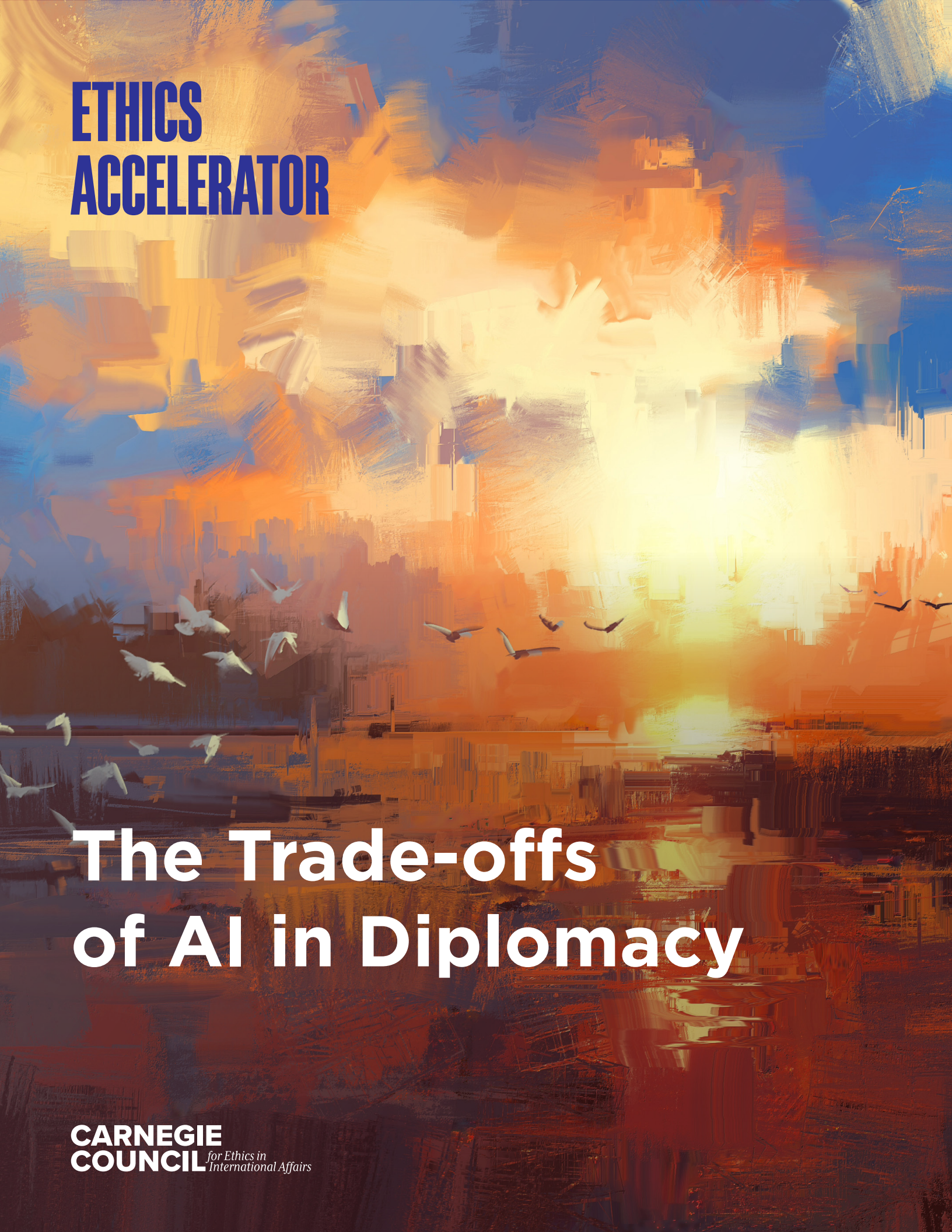
ETHICS
ACCELERATOR
The Trade-offs of AI in Diplomacy
CARNEGIE COUNCIL for Ethics in International Affairs

The Art of Diplomacy

Like any art, diplomacy requires skill, thought, and effort. In this particular art, much of that effort is dedicated to taking in, processing, and putting out a great deal of information.

Diplomatic entities must, for example, be able to summarize vast, complex bodies of text and data. This is so that they can inform their work with a detailed history of the issues and draw from their institutional memories. They must comprehend how their constituencies will be impacted by an issue, and know—deeply and accurately—their counterparties in any negotiation or conflict.

Diplomacy is also a generative art. Diplomats have to produce a wide range of media: speeches, communiqués, memoranda, simulations, policy ideations, position papers, resolutions, research reports, simulations, and educational materials. A lot of these media are rote repackagings of the same information, over and over. Some of them, however, are not.

Working in the international arena requires making difficult decisions. Whether it be the processing of visa applications to spur-of-the-moment choices in a live negotiation, these decisions call for smart prediction and inference in the face of uncertainties.

Finally, and not insignificantly, diplomacy involves a great deal of translation. Words must be conveyed from one language to another—often in real-time and under duress—without losing their meaning, their tone, or the innumerable fleeting subtleties that they contain.

A Right and an Obligation

Many believe that large language models (LLMs) could support, or fully take on, any number of the roles described above.

If these theories are right, it could be a good thing. Parties that are more efficient and effective might be more likely to succeed in diplomatic efforts, and will be able to better serve their policy goals and their constituents.

More efficient diplomatic organizations can also, ideally, better contribute to the ultimate goals of multilateralism: peace, stability, human rights, and social and environmental justice.

However, any application of large language models in diplomacy, as with their use in any other domain of life, will implicate trade-offs. These trade-offs vary according to how the LLM is used, the manner of its use, and the measures that are taken to adhere to ethical principles.

Some trade-offs are concrete and measurable, while others may build up silently and invisibly over time; they may be theoretical right up until the point that they have already caused an irreversible harm.

But one way or another, none of these trade-offs can be entirely dismissed.

The Trade-offs

Here is a list of those trade-offs:

A new vector of attack.
Each time a manual task is automated by computerized means, a new possibility of a cyberattack is created. Large language models are susceptible to a range of attacks that reduce their effectiveness and/or cause them to produce harmful outputs. Developing tailored AI for certain diplomatic roles will require states to compile and reproduce sensitive datasets—for example, a set of diplomatic cables—and these datasets can be hacked, too. The more widely an organization uses AI to do its work, and the more that it reduces the human role as a result, the greater the potential destructive effects of an attack. The possibility of an attack on AI can never be reduced to nil.

A loss of transparency, traceability, and accountability.
Large language models are inscrutable stochastic systems developed in closed environments, often by corporations that are unwilling to share information about their architecture. This makes it difficult to know how and why a system achieved a particular output. That, in turn, makes it difficult to trace the cause of—and hold the right people accountable for—harms that might arise from system outputs.

A rise in inequality.

Anybody with an Internet connection can use a commercially available chatbot. But making the most out of AI, especially in critical diplomatic tasks, will require tailored large language models and/or extensive skilled staffing. This could put such technologies beyond the reach of less resourced diplomatic parties. Furthermore, the technical foundations of LLMs and the datasets on which they are built over-represent the norms, traditions, languages, and cultures of the Global North while generally under- and mis-representing everyone else. Meanwhile, these systems' safety features are less effective against malicious attacks or use in Global South languages. As a result, the use of LLMs by wealthier states could exacerbate the North-South disparities that are widely agreed to be a major impediment to effective, sustainable, representative multilateral action.

An emergent unpredictability.

Large language models are complex systems that produce unpredictable outputs. The international stage is also a large complex system. When AI is used for diplomacy, these large complex systems will interact in complex ways. This could yield emergent effects; unplanned outcomes that are impossible to anticipate. Even when AI systems are seemingly used for non-critical tasks, these effects can rise to the surface quietly and incrementally, making an already turbulent global arena all the more volatile.

A skill, faded.

Many of the tasks that happen behind the scenes in diplomatic efforts—summarizing lengthy bodies of text, retrieving obscure pieces of information, translating routine comuniqués and other texts—might seem rote. Nevertheless, they require and reinforce important skills that go to the heart of effective diplomacy. An organization that leans heavily on AI may gain efficiencies in the short term, but they might lose certain key skills among their staff. In the long term, an organization that suffers a net loss in its human skill-base will be a less effective and reliable organization. In crises or other critical situations where the use of AI is either inadvisable or impractical, this skill fade could have a serious cost.

An in-human touch.

Ultimately, the art of diplomacy—like any art—relies on the human touch. Humans, not machines, are responsible for representing their constituents, for finding common ground with their counterparties, and striving toward the goals of diplomacy. Only a human can make a decision, large or small. Automating tasks in diplomacy reduces this human touch. While it may be true that a more efficient organization will be able to dedicate more of its human touch to the tasks where it matters most, it is difficult to define what tasks require that touch and which tasks do not. In some arenas, we may not miss the human touch until we've lost it entirely.

A State of Ethics

States and other parties involved in diplomacy have a right to explore the use of novel technologies that might make them more effective and efficient. But with this right comes an obligation to think deeply about the trade-offs involved in using those technologies.

Having ethical principles does not, on its own, manifest a state of ethics. Rather, ethical institutions deliberately and continuously study the trade-offs involved in their choices. Furthermore, ethics only thrives in the open; behind closed doors, it often wastes away.

In the diplomatic sphere, parties therefore have an obligation—both to their constituents and to their counterparties—to study the trade-offs that their use of LLMs entails. This document is intended as a basic guiding framework for that work.

We urge states to engage in this thinking transparently and be open with their constituents and their counterparties about their awareness and acceptance of the trade-offs of LLMs in their work.

Then, and only then, can the technology's potential benefits be ethically reaped.

Carnegie Ethics Accelerator

We live in a moment when new ethical questions are emerging at an exponential rate. Society faces significant challenges in the realm of international affairs as new technologies are developed, deployed, and co-opted with haste by actors who view ethics as an encumbrance rather than a requisite. In response, Carnegie Council launched the **Ethics Accelerator**, a new kind of incubator that seeks to address technology ethics issues in a manner that matches the pace at which new technologies emerge and proliferate.

This communiqué was developed as part of an Ethics Accelerator convening that assembled experts from across the legal, technology, diplomatic, academic, and NGO communities, and was held under Chatham House Rule. The Carnegie Ethics Accelerator is generously supported by the Patrick J. McGovern Foundation.

Carnegie Council for Ethics in International Affairs

At **Carnegie Council**, we believe that ethics is at the heart of our greatest global challenges and that by working to empower ethics, we can discover common values and interests that lead to a better future.

Founded by Andrew Carnegie over a century ago, we set the global ethical agenda and work for an ethical future by **identifying** current and future critical ethical issues, **convening** leading experts, **producing** agenda-setting resources, and **catalyzing** the creation of ethical solutions to global problems. **Join us in using the power of ethics to build a better world.** Carnegie Council is a nonprofit 501(c)(3) institution.

Contact:
info@cceia.org
212-838-4120
170 East 64th Street
New York, NY 10065