Schopmans, Hendrik; Cupać, Jelena

Engines of Patriarchy: Ethical Artificial Intelligence in Times of Illiberal Backlash Politics

Ethics & International Affairs

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Schopmans, Hendrik; Cupać, Jelena (2021) : Engines of Patriarchy: Ethical Artificial Intelligence in Times of Illiberal Backlash Politics, Ethics & International Affairs, ISSN 1747-7093, Cambridge University Press, Cambridge, Vol. 35, Iss. 3, pp. 329-342, https://doi.org/10.1017/S0892679421000356

ESSAY

Engines of Patriarchy: Ethical Artificial Intelligence in Times of Illiberal Backlash Politics

Hendrik Schopmans and Jelena Cupać

In October 2020, discontent slowly took root at the Council of the European Union. In the months prior, the council had worked hard to develop a European position on artificial intelligence (AI), one that would firmly reflect the EU's commitment to fundamental rights. As its members were preparing to adopt conclusions on the matter, however, it became apparent that no consensus could be reached. Opposition came from one member—Poland—that fixated on the mention of "gender equality" in the draft conclusions. The Polish representatives argued that because the term "gender" did not appear in the EU Charter of Fundamental Rights, there was no need to mention it in the context of AI.[1] Other council members disagreed vehemently, yet they proved unable to dissuade Poland from its stance. Eventually, the German council presidency concluded that efforts to forge a consensus had failed. Left without any alternatives, it adopted presidency conclusions, which did not require member states' unanimous agreement and merely expressed the presidency's position on the matter.

For observers of the emerging AI policy landscape, the preceding episode stands out as a curious anomaly. Until recently, finding common ground on the high-

Hendrik Schopmans, WZB Berlin Social Science Center, Berlin, Germany; Free University of Berlin, Berlin, Germany (hendrik.schopmans@wzb.eu).

Jelena Cupać, WZB Berlin Social Science Center, Berlin, Germany (jelena.cupac@wzb.eu).

Ethics & International Affairs, 35, no. 3 (2021), pp. 329–342.

doi:10.1017/S0892679421000356

level principles and values that should guide future AI governance efforts had been largely uncontroversial. In fact, over the past five years, actors across the world have responded to mounting concerns over the potential downsides of progress in AI—such as the proliferation of discriminatory algorithms, the specter of automated mass surveillance, and the existential threat posed by a future superintelligence—by unleashing a downright wave of documents and principles on AI ethics.[2] In these documents, governments, corporations, researchers—and even the Pope—have presented themselves as vocal advocates of "responsible," "trustworthy," and "human-centered" AI.[3] Crucially, many of them have embraced similar tenets, pledging to develop and deploy AI in accordance with principles of transparency, justice and fairness, and privacy.[4] Given the nonbinding nature of these documents, the prevalence of consensus may come as no surprise, as it likely masks diverging positions on the exact meaning of these principles. At the same time, regional and international organizations (IOs) have begun to build on this high-level convergence to develop more concrete international rules and standards.[5] Against this background, Poland's refusal to consent to a largely symbolic statement on AI, prompted only by the inclusion of a specific term, stands out.

In this essay, we argue that what seems like an isolated incident in the context of AI was, in fact, a manifestation of a broader trend in international politics: the illiberal backlash against the liberal international order. In recent years, a diverse constellation of actors, positioned within and outside of liberal democracies, has increasingly challenged the values and norms that have dominated international politics since the end of the Cold War.[6] Although the normative targets of this backlash have varied—ranging from human rights and gender equality to environmental protection—the contesters have converged in one central critique: their rejection of an international order that is imbued with principles of political liberalism and that, above all, seeks to promote and protect individual rights.[7] In an effort to counter this institutionalized liberal bias in the order's social purpose, the governments, parties, and nongovernmental organizations (NGOs) driving the recent "tide of illiberalism" have promoted a reactionary agenda.[8] Their goals coalesce around the global promotion of traditional value systems, be it a more prominent role for state-organized religion, heterosexual family values, or the protection of national and cultural identities against external influences.

We propose here that this wider illiberal backlash is likely to have substantial consequences for the future of AI governance. First of all, illiberal contesters

have coordinated to openly attack—and overturn—some of the liberal norms underpinning the current international order. As they contest progressive causes across issue areas, we expect them to also challenge the validity of certain norms in the context of AI. Here, the real danger lies in the strategies that illiberal actors employ in the process. Rather than engaging in innocuous language games, we show that illiberal contesters are increasingly willing to *spoil* urgent international negotiations if they are unable to get their way on single issues. We thus expect them to be prepared to veto AI legislation as a whole if certain norms, rules, or even singular terms are not watered down or eliminated according to their wishes.

To illustrate this argument, we focus on gender equality as a norm that has been at the heart of the illiberal backlash. We show that just as the fight against gender-based discrimination has gained momentum within the AI research community, gender equality norms have come under sustained attack in international institutions. Our analysis reveals that across various issue areas, illiberal regimes, regional groupings, and antifeminist NGOs have formed coalitions of convenience to block and drive back international efforts at strengthening women's rights. We conclude that backlash politics not only threatens the entrenchment of progressive norms in AI legislation but also jeopardizes the crucial move from high-level principles to robust international rules for AI more generally.

Our appeal to take illiberal backlash politics seriously is primarily directed at scholars and practitioners concerned with AI ethics and governance. So far, discussions on the future prospects of global AI governance have only paid limited attention to the global political context in which it unfolds. The one exception is security-centered accounts that predominantly examine the global politics of AI through the prism of the "AI arms race."[9] This perspective holds that rival states and corporations are prone to think of AI development as a zero-sum game. As they strive for leadership in a transformative technology domain, competitors view governance and safety considerations as an unnecessary obstacle to getting ahead in the race.[10] Ethical commitments amount to little more than "ethics washing," a form of cheap talk that is not backed up with meaningful action.[11] Although the current focus on competitive dynamics and ethics washing is justified, we argue here that a potentially more significant threat to the protection of fundamental rights in the age of AI is looming in the assembly halls of the United Nations and other IOs.

Modelling Patriarchy: The Creation of Biased Artificial Intelligence

The risk of designing AI systems that reproduce societal biases is a prominent driver of current debates over global rules for AI. Many concerns over bias are related to the data dependency of contemporary AI systems; in particular, machine-learning algorithms referred to as deep neural networks. Because neural networks infer patterns from massive amounts of already *existing* data—be it numbers, words, or images—they are inherently conservative. Any bias that characterizes a society—and hence the data that describe it—will be reproduced in the respective AI system. From seat belt design to symptoms of cardiovascular diseases, the data describing many aspects of society have long been shaped by the needs, wants, and beliefs of white men.[12] Such data are bound to put everyone else at risk, simply because there is less data available on other groups and because the data often reflect historical patterns of injustice.

At the same time, bad data are not the only culprit. The relative autonomy of AI systems notwithstanding, they are ultimately products of human design. This means that both the individual biases of developers and the structural biases permeating the field of AI at large drive choices regarding which data are selected, how they are labeled, and to what end they are used. Such decisions, too, are consequential for whether a diagnostic tool detects a fatal disease, whom a predictive-policing software identifies as a criminal suspect, and whether controversial applications, such as emotion recognition systems, ever see the light of day.

Biased AI is not merely a hypothetical risk. As an increasingly vocal group of researchers and activists have pointed out, many AI systems already reproduce societal patterns of gender discrimination. In 2016, a Boston-based research group set out to create an "analogy generator" using Word2Vec, a tool that captures relationships between words and that had been trained on a corpus of Google News texts. When asked to fill in missing words in analogies, the model reproduced sexist stereotypes it had learned from the data. Most notably, its output stated that "Man is to computer programmer as woman is to homemaker" and "A father is to a doctor as a mother is to a nurse."[13] Two years later, technology giant Amazon was forced to abandon an experimental hiring tool that ranked prospective job candidates by autonomously screening their applications. Trained on the resumes of former applicants, who were predominantly male, the system had taught itself to penalize resumes that featured the word "women."[14]

The pervasiveness of gender bias in AI systems is not limited to the realm of language. In 2018, researchers Joy Buolamwini and Timnit Gebru demonstrated that commercial facial recognition software was substantially less accurate in identifying dark-skinned females than light-skinned males. A system developed by IBM, for instance, produced an error rate of 34.7 percent for dark-skinned women compared to 0.3 percent for light-skinned males.[15] Considering the widespread adoption of facial recognition software in security apparatuses—from border controls to investigations of political assemblies—the potential for fatal decisions made by biased AI systems has become very real.[16]

In the struggle for fairness in AI systems, researchers and activists have been largely left to their own devices. Some have made notable achievements in mobilizing support, pressuring corporations, and removing bias by technological means, for example, by developing "debiasing" techniques and constructing more representative data sets. Nonetheless, researchers still face an uphill battle to change the larger power structures that permit and sustain bias. Most research pertaining to bias has come from women, who are starkly underrepresented in the field of AI.[17] According to the World Economic Forum, in 2018 only 22 percent of AI professionals were female.[18] At Google, one of the biggest players in AI development, just 10 percent of the research staff are women, and only 2.5 percent of the workforce are Black. In December 2020, the technology giant ignited a renewed debate about diversity in AI when it fired Timnit Gebru, co-lead of the company's ethical AI team, over a research paper she had coauthored. In the paper, Gebru, an outspoken advocate on issues of diversity, and her fellow researchers had pointed out the various harms involved in the use of large language models—including their propensity to reproduce bias embedded in language.[19] Overall, in a field that continues to be "extremely white, affluent, technically oriented, and male,"[20] pledges to ensure diversity and fairness have often amounted to little more than rhetoric.

Against this backdrop, current efforts to create global standards for AI are particularly significant. International organizations are important fora in which the norms that should guide AI development are articulated and supplemented with binding rules, standards, and certification procedures. IOs, therefore, have the power to ensure that the protection of gender equality norms in AI does not only depend on the bottom-up efforts of researchers or the goodwill of technology corporations. Regulatory bodies such as the European Commission could move toward penalizing the use of discriminatory AI systems and shift resources

toward increasing diversity in AI research. This, however, is where the global political context matters. The promotion of gender equality norms for AI by international institutions presupposes international agreement on gender equality as a value to be cherished in the first place. And on the global stage, the backlash against it is growing.

Countering "Gender Ideology": The Global Antifeminist Backlash

As part of the broader illiberal attack on liberal international institutions, emboldened authoritarian regimes, populists, ultraconservative governments, and antifeminist NGOs have been particularly adamant about countering progressive gender norms and women's rights. Their heightened activity in halting and forcing back these norms, coupled with their work on instituting so-called pro-family and pro-life causes, has prompted many to speak of a full-scale antifeminist backlash.[21] The backlash has gained particular traction in Europe, where large-scale protests against the Council of Europe's 2011 Convention on Preventing and Combating Violence against Women and Domestic Violence, also known as the Istanbul Convention, have taken place. The convention has been attacked for its alleged attempt to institute "gender ideology"—a tenet that, according to its contesters, corrodes traditional family and gender roles.[22]

The United Nations is also experiencing a surge in conservative and antifeminist activism. Groups pushing against women's rights have been a fixture in the UN since the early 1990s. However, only in the past decade have they grown to a size that allows them to do more than merely tinker on the sidelines of important meetings. In addition to an increasing number of antifeminist NGOs, many Catholic, Islamic, and post-Soviet states; the United States under the Trump administration; and the Vatican have actively contested women's rights and gender norms as part of their UN agenda.[23] Cognizant of their shared normative program, these diverse actors have coordinated to form coalitions of convenience—causing several human rights organizations to sound the alarm over an increasingly organized "unholy alliance."[24]

For instance, in 2015 a number of UN member states founded the Group of the Friends of the Family, while antifeminist NGOs also jointly operate through both the UN Family Rights Caucus and Civil Society for the Family. The World Congress of Families (WCF), organized by U.S.-based organizations representing

the Christian Right, has taken on a particularly prominent role in bringing together pro-family governments and societal actors from across the globe. Since 2012, the WCF has held annual conferences designed, among other things, to strategize on how to best insert the language of the "natural family" and "traditional values" into UN documents.[25] Two of these conferences were held in Hungary and Italy, with Hungarian prime minister Viktor Orbán and then–Italian deputy prime minister Matteo Salvini addressing the audience. On various occasions, actors contesting women's rights in the UN have also lobbied for and successfully secured the support of regional groupings such as the Organisation for Islamic Cooperation, the League of Arab States, the Africa Group at the UN, and the G-77.[26] Varying the composition of these coalitions of convenience across issue areas, contesters of women's rights have successfully joined forces to secure victories in various UN fora.[27]

The strategy that the contesters consistently deploy is seemingly simple. They first identify the progressive language in international documents and then endeavor to water it down or delete it entirely. Besides the word "gender," they target terms such as "sexual and reproductive health and rights," "comprehensive sexual education," and "women and girls in all their diversity," among others. One widely publicized episode from the United Nations Security Council (UNSC) illustrates this approach well. In early 2019, the UNSC prepared to adopt a ninth resolution in its Women, Peace and Security agenda. The resolution was relatively uncontroversial: it sought to bolster domestic and international efforts to combat sexual violence in conflicts. Still, it attracted considerable controversy. Contention centered on the phrase "sexual and reproductive health." The U.S. representatives interpreted the phrase as a euphemism for abortion and, threatening to veto the resolution, demanded that it be deleted from the document. The threat worked. The remaining UNSC members concluded that it was better to adopt a watered-down version of the resolution than not to adopt one at all.

As this example shows, the language strategies used by contesters of women's rights and progressive gender norms are underpinned by a particular kind of politics, at the root of which is the high moral status they give to their pro-life and pro-traditional family beliefs. In very basic terms, they define "the life of the unborn child" and "the natural family" as their most significant values—values that allow for little compromise.[28] If a document appears to threaten these values, even just by using the term "gender," the contesters show readiness to defend them, to the detriment of any other concern the document might be tackling.

As the UNSC example shows, even a relatively uncontroversial humanitarian initiative can be undermined if it indicates a possible international legal recognition of abortion. This does not mean that conservatives and antifeminists regard other global concerns as entirely unimportant. Rather, it means that they have ordered their moral preferences in such a way that they find it justified to spoil *any* international initiative they consider a threat to their values.

And indeed, we observe that contesters of women's rights employ this spoiling approach across many different issue areas in which they identify a progressive challenge, from same-sex relationships and affirmative action to disability rights.[29] Most recently, the strategy was observable in international deliberations on the COVID-19 pandemic. For most countries, the pandemic's health and economic costs have been overwhelming, making robust multilateral responses ever more necessary. Yet, women's rights contesters have continued to browse UN draft documents, searching for threats to their values. By objecting to the phrase "sexual and reproductive health," antifeminist NGOs attacked the UN's COVID-19 funding plan, while several states issued amendments to the General Assembly's 2020 draft resolutions on providing health care to women and girls during the pandemic.[30]

In summary, over the past decade, contesters of women's rights have not only steadily increased their presence in international organizations; they have also increasingly coordinated their activities. The most visible result of this coordination is the consistent deployment of a common contestation strategy—one that is characterized by the willingness to throw a wrench in any international initiative that, even marginally, seems to further progressive ideas on gender and women's rights.

The Backlash Extends: Contesting Gender Equality in AI Governance

Considering how relentlessly the contesters have extended their fight into new issue areas, there is little reason to believe that AI governance should be an exception. In fact, the increased awareness of gender bias in AI systems has caught the attention of global regulators at the same time as numerous conservative and antifeminist actors have firmly positioned themselves within key institutions of global governance.

Taking these concurrent developments into account, the incident that opened this essay appears in a new light—no longer as an isolated episode, but as the first instance of inevitable friction. First of all, Poland's government matches the profile of a conservative and increasingly illiberal regime. As testified by the government's crackdown on LGBTQI and women's rights domestically, it has embraced a reactionary agenda and taken an increasingly combative stance on progressive causes. Yet even more telling is the pattern of behavior it exhibited at the Council of the European Union. Its actions are a prime example of the spoiling strategy we describe above: In this case, the contester, Poland, preferred to undercut the agreement on a joint agenda over allowing it to be adopted with a progressive understanding of gender—despite the fact that the issue of gender equality constituted only one part of a much broader agenda. Only a couple of months later, the country—this time joined by another illiberal contester, Hungary—opposed another EU initiative on similar grounds, obstructing a plan to "promote gender equality and women's empowerment" in EU foreign policy.[31]

As international bodies move from ethical commitments to legal frameworks for AI, we can expect such backlash against gender equality norms to intensify. The implications are serious: By threatening to veto not only funding plans and normative declarations but entire legislative proposals, contesters may bully their way to watering down, or outright eliminating, progressive gender norms in AI governance. At the same time, insistence on the inclusion of these norms may lead to the breakdown of negotiations altogether. As the past behaviors of those driving the antifeminist backlash have shown, their threats are rarely empty.

An extension of the challenge to the domain of AI becomes even more likely when we consider the sheer number of IOs that have picked up on the issue of gender discrimination in AI. A draft recommendation by UNESCO, meant to initiate the first global standard-setting instrument on AI ethics, strongly emphasized the issue of gender equality. Similarly, the European Commission's white paper on AI, a precursor to its planned legal framework, proposes to address the issue of gender-based discrimination through the use of more representative data sets. Overall, we expect that the more prominently gender equality features in international legal frameworks for AI, the more resistance it will draw. As in other issue areas, illiberal governments will likely be joined by conservative and antifeminist groups once they realize AI is a decisive new front in the global struggle against liberal norms and values.

Conclusion

This essay has brought together two political developments that have so far been treated in isolation: the emerging efforts to create international legal frameworks for AI and the contestation of the liberal international order. We have argued that the global antifeminist backlash presents a serious obstacle to embedding gender equality norms in global AI governance. Importantly, we observe that this backlash also threatens AI regulation efforts more generally. An increasingly coordinated group of actors have imbued their pro-life and pro-traditional family beliefs with a moral status of the highest order, thus precluding compromise with more progressive agendas. If an instance of AI regulation contains even a trace of "gender ideology," these actors will not hesitate to veto it—regardless of the urgency for transforming ethical principles for AI into more implementable norms, rules, and standards.

At this point, we recognize that our argument may read as a warning of what is yet to come; still, we do not believe it to be purely speculative. We draw our conclusions from well-observed antifeminist trends in other issue areas. In the face of the harm that biased AI systems can inflict upon vulnerable groups, we believe that identifying the potential obstacles to the regulation of discriminatory AI systems at an early stage is not only desirable, but vital. In light of this, we argue that Poland's recent behavior at the Council of the European Union likely marks merely the start of the collision between the growing antifeminist backlash and the global push for AI legislation.

It is worth widening our view here to other areas of AI ethics that may also fall victim to illiberal backlash politics. Like gender equality, other human rights norms that prohibit discrimination—for example, along racial lines—may also come under attack. In liberal democracies, backlash politics is underpinned by the revival of national and ethnic identities, such as the resurgence of white supremacy in the United States and parts of Western Europe. Protecting minorities from discrimination by automated decision-making systems may, therefore, be placed low on the list of political priorities, and, even more troublingly, governments may intentionally employ these technologies against minorities. A blueprint for doing so already exists. China, for instance, is increasingly drawing on AI-based surveillance tools in its violent suppression of the Uighur minority.[32] In December 2020, Chinese technology giant Alibaba was exposed for developing facial recognition technology that could specifically identify Uighurs.[33] In the

absence of any international legal or regulatory pressures, technology corporations may increasingly become complicit in these efforts.

To conclude, while the emerging, high-level consensus on "ethical" and "human-centered" AI may seem promising, it hides fundamentally different conceptions of whose ethics should apply to which humans. To be clear, we believe that an inclusive, critical debate about the current dominance of liberal conceptions of ethics in AI is certainly necessary. Yet, the uncompromising behavior displayed by illiberal contesters thus far has engendered intractable polarization rather than encouraged careful deliberation. Most importantly, it has increased the risk of harmful outcomes. It might lead to AI legislation that turns a blind eye to biased AI systems or, as we have shown, prevents much needed rules for AI from being adopted altogether. In view of this, scholars and practitioners concerned with AI ethics need to take seriously illiberal forces and their attempts to shape the still contested meaning of ethical AI. In the first instance, they need to pay closer attention to international institutions as key spaces where AI norms preserving fundamental rights are contested—and where they must be proactively defended. Most importantly, scholars and practitioners should be sensitized to spoiling strategies in order to recognize them and call them out as self-interested attempts to hold negotiations hostage. Understanding the potential intentions behind the actions and words of illiberal actors constitutes the first step toward this end—and toward ensuring that AI will truly serve humanity as a whole.

NOTES

[1] The full Presidency Conclusions on the charter of fundamental rights in the context of AI, including a short note on Poland's opposition, can be accessed at "Artificial Intelligence: Presidency Issues Conclusions on Ensuring Respect for Fundamental Rights," European Council/Council of the European Union, October 21, 2020, www.consilium.europa.eu/de/press/press-releases/2020/10/21/artificial-intelligence-presidency-issues-conclusions-on-ensuring-respect-for-fundamental-rights/#.

[2] Daniel Schiff, Jason Borenstein, Justin Biddle, and Kelly Laas, "AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection," *IEEE Transactions on Technology and Society* 2, no. 1 (March 2021), pp. 31–42; and Merve Hickok, "Lessons Learned from AI Ethics Principles for Future Actions," *AI and Ethics* 1 (February 2021), pp. 41–47.

[3] See, for example, Jen Copestake, "AI ethics backed by Pope and tech giants in new plan," BBC News, February 20, 2020, www.bbc.com/news/technology-51673296; Nathalie A. Smuha, "The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence," *Computer Law Review International* 20, no. 4 (August 2019), pp. 97–106; and Natasha Crampton, "The Building Blocks of Microsoft's Responsible AI Program," *Microsoft on the Issues* (blog), January 19, 2021, blogs.microsoft.com/on-the-issues/2021/01/19/microsoft-responsible-ai-program/.

[4] Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1, no. 9 (September 2019), pp. 389–99; and Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI" (HLS white paper, Berkman Klein Center for Internet & Society, Harvard University, 2020).

[5] For instance, following consultations on ethical principles by a high-level expert group, the European Commission released its first draft of the Artificial Intelligence Act in April 2021. UNESCO, meanwhile,

is currently developing the first global standard–setting instrument on the ethics of artificial intelligence.

[6] For a conceptualization of backlash politics, see Karen J. Alter and Michael Zürn, "Conceptualising Backlash Politics: Introduction to a Special Issue on Backlash Politics in Comparison," *British Journal of Politics and International Relations* 22, no. 4 (November 2020), pp. 563–84.

[7] Tanja A. Börzel and Michael Zürn, "Contestations of the Liberal International Order: From Liberal Multilateralism to Postnational Liberalism," in "Challenges to the Liberal International Order: International Organization at 75," special issue 2, *International Organization* 75 (Spring 2021), pp. 282–305, www.cambridge.org/core/journals/international-organization/article/contestations-of-the-liberal-international-order-from-liberal-multilateralism-to-postnational-liberalism/7CE3FD0F629D18BE45EB9C7AC70954AA.

[8] Alexander Cooley and Daniel H. Nexon, "The Illiberal Tide: Why the International Order Is Tilting toward Autocracy," *Foreign Affairs*, March 26, 2021, www.foreignaffairs.com/articles/united-states/2021-03-26/illiberal-tide?utm_medium=promo_email&utm_source=lo_flows&utm_campaign=registered_user_welcome&utm_term=email_1&utm_content=20210512.

[9] On the AI arms race, see Paul Scharre, "Killer Apps: The Real Dangers of an AI Arms Race," *Foreign Affairs* 98, no. 3 (May/June 2019), pp. 135–44; Peter Asaro, "What Is an 'Artificial Intelligence Arms Race' Anyway?," *I/S: A Journal of Law and Policy for the Information Society* 15, nos. 1–2 (2019), pp. 45–64; Benjamin M. Jensen, Christopher Whyte, and Scott Cuomo, "Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence," *International Studies Review* 22, no. 3 (September 2020), pp. 526–50; and Amandeep Singh Gill, "Artificial Intelligence and International Security: The Long View," *Ethics & International Affairs* 33, no. 2 (Summer 2019), pp. 169–79.

[10] Stephen Cave and Seán S. ÓhÉigeartaigh, "An AI Race for Strategic Advantage: Rhetoric and Risks" (conference presentation, AIES '18: AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, February 2–3, 2018), in *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New York: Association for Computing Machinery, December 2018), pp. 36–40.

[11] On ethics washing, see, e.g., Luciano Floridi, "Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical," *Philosophy & Technology* 32, no. 2 (June 2019), pp. 185–93; and Anaïs Rességuier and Rowena Rodrigues, "AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics," *Big Data & Society* 7, no. 2 (July 2020).

[12] Caroline Criado Perez, *Invisible Women: Exposing Data Bias in a World Designed for Men*, 1st ed. (London: Chatto & Windus, 2019).

[13] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam T. Kalai, "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings," *Advances in Neural Information Processing Systems* 29 (2016), pp. 4349–57, arxiv.org/abs/1607.06520.

[14] Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women," Reuters, October 10, 2018, www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[15] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" (conference presentation, ACM [Association for Computing Machinery] Conference on Fairness, Accountability, and Transparency, New York, February 24, 2018), in *Proceedings of Machine Learning Research* 81 (2018), pp. 1–15, proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[16] In late 2020, for instance, the *New York Times* reported three cases of black men being wrongfully arrested in the United States based on a false match by facial recognition technology. See Kashmir Hill, "Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match," Technology, *New York Times*, updated January 6, 2021, www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html.

[17] Susan Leavy, "Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning" (conference presentation, IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering, Gothenburg, Sweden, May 27–June 3, 2018), in *Proceedings: 2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE 2018)* (New York: Association for Computing Machinery), pp. 14–16.

[18] "Assessing Gender Gaps in Artificial Intelligence," in *The Global Gender Gap Report 2018* (Cologny, Switzerland: World Economic Forum, 2018), pp. 29–31, reports.weforum.org/global-gender-gap-report-2018/assessing-gender-gaps-in-artificial-intelligence/.

[19] Alex Hanna and Meredith Whittaker, "Timnit Gebru's Exit from Google Exposes a Crisis in AI," *WIRED*, December 31, 2020, www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai/; and Karen Hao, "We Read the Paper That Forced Timnit Gebru Out of Google. Here's What It Says," *MIT Technology Review*, December 4, 2020, www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/.

[20] Sarah Myers West, Meredith Whittaker, and Kate Crawford, *Discriminating Systems: Gender, Race and Power in AI* (New York: AI Now Institute, New York University, 2019), p. 6, ainowinstitute.org/discriminatingsystems.pdf.

[21] Jelena Cupać and Irem Ebetürk, "The Personal Is Global Political: The Antifeminist Backlash in the United Nations," *British Journal of Politics and International Relations* 22, no. 4 (November 2020), pp. 702–14; Anne Marie Goetz, "The New Cold War on Women's Rights?," United Nations Research Institute for Social Development, June 22, 2015; Judith Butler, "The Backlash against 'Gender Ideology' Must Stop," *New Statesman*, January 21, 2019; Barbara Crossette, "At the UN, Twenty Years of Backlash to 'Women's Rights Are Human Rights,'" *Nation*, March 5, 2013; Neil Datta, "'Restoring the Natural Order': The Religious Extremists' Vision to Mobilize European Societies against Human Rights on Sexuality and Reproduction" (Brussels: European Parliamentary Forum on Population & Development, 2018); and Isabel Marler and Naureen Shameem, "Human Rights Are under Attack by an Ultra-Conservative Agenda," Association for Women's Rights in Development, December 7, 2016.

[22] For instance, in a letter from 2018, addressed to Secretary General of the Council of Europe Thorbjørn Jagland, 333 NGOs asked for a revision of the Istanbul Convention in order to remove its "ideological" components (For the full text, see Cristiana Scoppa, "WAVE (Women against violence Europe) chiede di contrastare l'azione reazionaria e di destra di quelle organizzazioni che stanno attaccando la Convenzione di Istanbul," *Womenews*, April 14, 2018, http://www.womenews.net/2018/04/14/wave-women-against-violence-europe-chiede-di-contrastare-lazione-reazionaria-e-di-destra-di-quelle-organizzazioni-che-stanno-attaccando-la-convenzione-di-istanbul/.

[23] Jelena Cupać and Irem Ebetürk, "Backlash Advocacy and NGO Polarization over Women's Rights in the United Nations," *International Affairs* 97, no. 4 (July 2021), pp. 1183–1201.

[24] Naureen Shameem, *Rights at Risk: Observatory on the Universality of Rights; Trends Report 2017* (Toronto: Association for Women's Rights in Development, 2017); and "Vatican in Unholy Alliance at United Nations," International Trade Union Confederation, March 8, 2013, www.ituc-csi.org/vatican-in-unholy-alliance-at?lang=en.

[25] Human Rights Campaign Foundation, "Exposed: The World Congress of Families; An American Organization Exporting Hate" (Washington, D.C.: Human Rights Campaign Foundation, updated June 2015), assets2.hrc.org/files/assets/resources/WorldCongressOfFamilies.pdf.

[26] Shameem, *Rights at Risk*; for the G-77, see Françoise Girard, "Taking ICPD beyond 2015: Negotiating Sexual and Reproductive Rights in the Next Development Agenda," *Global Public Health* 9, no. 6 (July 2014), pp. 607–19.

[27] These fora include the Security Council, the Human Rights Council, and the Commission on the Status of Women.

[28] See, for instance, Allan C. Carlson and Paul T. Mero, *The Natural Family: A Manifesto* (Dallas: Spence, 2007); and Sharon Slater, *Stand for the Family: Alarming Evidence and Firsthand Accounts from the Front Lines of the Battle; a Call to Responsible Citizens Everywhere* (Mesa, Ariz.: Inglestone, 2009).

[29] Anne Marie Goetz, "The New Competition in Multilateral Norm-Setting: Transnational Feminists & The Illiberal Backlash," *Dædalus* 149, no. 1 (Winter 2020), pp. 160–79; and Shameem, *Rights at Risk*.

[30] See Stefano Gennarini, "Abortion 'Essential' in UN's $2B COVID-19 Funding Plan," Center for Family & Human Rights, April 2, 2020, c-fam.org/friday_fax/abortion-essential-in-uns-2b-covid-19-funding-plan/; and Julian Borger, "Trump Administration in 'Staggering' Isolation at UN on Health Issues," *Guardian*, November 19, 2020, www.theguardian.com/world/2020/nov/19/trump-administration-in-staggering-isolation-at-un-on-health-issues.

[31] Quoted in Hans von der Burchard, "EU's Foreign Policy Gender Plan Faces Resistance from Poland and Hungary: Two Countries Oppose EU Plan to Bolster Women's, Girls' and LGBTQI Rights," *POLITICO*, November 25, 2020, www.politico.eu/article/eus-gender-equality-push-for-external-relations-faces-trouble-from-the-inside/.

[32] Ross Andersen, "The Panopticon Is Already Here: Xi Jinping Is Using Artificial Intelligence to Enhance His Government's Totalitarian Control—and He's Exporting This Technology to Regimes around the Globe," *Atlantic*, September 2020, www.theatlantic.com/magazine/archive/2020/09/china-ai-surveillance/614197/?utm_source=twitter&utm_medium=social&utm_campaign-=share.

[33] Reuters, "Alibaba Facial Recognition Tech Specifically Picks Out Uighur Minority," Reuters, December 17, 2020, www.reuters.com/article/us-alibaba-surveillance-idUSKBN28RoIR.

Abstract: In recent years, concerns over the risks posed by artificial intelligence (AI) have mounted. In response, international organizations (IOs) have begun to translate the emerging consensus on

the need for ethical AI into concrete international rules and standards. While the path toward effective AI governance faces many challenges, this essay shifts attention to an obstacle that has received little attention so far: the growing illiberal backlash in IOs. Prompted by Poland's recent rejection of a European position on AI due to the document's mention of "gender equality," we argue that Poland followed a strategy that illiberal actors now regularly employ in IOs. To combat gender norms and women's rights across issue areas, illiberal contesters first identify the progressive language in international documents and then threaten to veto those documents—unless such language is watered down or removed. This spoiling strategy, we argue, may not only lead to the compromising of fundamental human rights norms but may also prevent much needed rules for AI from being adopted altogether. Against this background, we urge scholars and practitioners concerned with AI ethics to pay closer attention to illiberal backlash politics. IOs are emerging as spaces where progressive AI rules and standards are increasingly contested—and where they need to be defended to safeguard fundamental rights in an age of rapid technological change.

Keywords: artificial intelligence, ethics, backlash, gender equality, liberal international order, gender bias