



ETC

ETHICAL TECH COLAB · NYU CENTER FOR GLOBAL AFFAIRS · RESEARCH PAPER

WHAT IS ETHICAL AI?

ETHICS, ETHICAL TECHNOLOGY, AND ETHICAL INTERNATIONAL RELATIONS FOR THE AGE OF INTELLIGENT MACHINES

The issue is power. Ethics is the answer. This paper traces ethics from the earliest civilizations through international affairs and human rights law to the responsible AI movement, the humanitarian sector, and the United Nations system, and closes with the motivation and research philosophy of the Ethical Tech CoLab.

Authors

Ethical Tech CoLab · NYU School of Professional Studies Center for Global Affairs
in collaboration with Microsoft

Master's Research · M.S. in Global Affairs

July 2026 · <https://ethical-tech-colab.github.io/website/> · github.com/Ethical-Tech-CoLab

What is ethics, and why did humanity invent it? What makes technology ethical, international relations ethical, and artificial intelligence ethical? This paper answers those questions in that order, on the conviction that the issue at the heart of artificial intelligence is power and that ethics is humanity's oldest and most tested answer to the problem of governing power. It traces ethics from its origins in the earliest civilizations through international affairs and human rights law to the responsible AI movement, the humanitarian sector, and the United Nations system, and closes with the motivation and research philosophy of the Ethical Tech CoLab.

The Ethical Tech CoLab is a research initiative of the NYU School of Professional Studies Center for Global Affairs, conducted in collaboration with Microsoft. Views and findings are those of the researchers and do not represent the official positions of New York University, Microsoft, or any partner institution. External programs cited are referenced as evidence, not as CoLab partnerships.

Contents

1. Executive Summary	4
2. Humanity's Relationship with Power	5
2.1 The Paradox of Capability	5
2.2 Power in Social Thought	5
2.3 The Recurring Pattern and the New Concentration of Power.....	6
3. What Is Ethics?	7
3.1 Definition and Deliberation	7
3.2 Ethics and Morality	7
3.3 Ethics, Law, and Power	8
4. Why Humanity Created Ethics	8
4.1 Cooperation and Survival	8
4.2 The Earliest Ethical Systems	9
4.3 Ethics as Institutional Infrastructure	9
5. From Custom to Universal Ethics: The Expansion of Moral Concern	10
5.1 The Expanding Circle.....	10
5.2 The Universal Declaration.....	10
5.3 Universalism Without Uniformity	11
6. What Is Ethical International Relations? Realism, Idealism, and Ethical Realism	11
6.1 Realism and Idealism	11
6.2 Ethical Realism.....	12
6.3 Ethical Deliberation as Method.....	12
7. What Is Ethical Tech?	13
7.1 Artifacts Have Politics.....	13
7.2 Value Sensitive Design and Responsible Innovation.....	13
7.3 Ethical Tech as Practice.....	14
8. What Is Ethical AI?	14
8.1 A Definition.....	14
8.2 From Principles to Governance	14
8.3 Accuracy Is Not Ethics.....	15
9. The Rise of Responsible AI.....	15
9.1 The Age of Principles.....	15
9.2 Ten Landmark Frameworks.....	16
9.3 The Reckoning.....	18
9.4 Operationalization and the Safety Turn.....	18
10. The Historical Evolution of AI in Humanitarian Work	19
10.1 Crowdsourcing Origins	19
10.2 Institutionalization and Prediction	20
10.3 Identification and Biometrics	20
10.4 The Generative Era	21
11. Why Humanitarian AI Is Fundamentally Different from Commercial AI	21
11.1 Principles Under Pressure.....	21
11.2 Consent and Power.....	22
11.3 Humanitarian Experimentation.....	22
11.4 Accountability and the Cost of Failure	22
12. The International Affairs Context.....	23
12.1 Three Digital Empires.....	23
12.2 The EU AI Act.....	24
12.3 The United States: A Different Theory of Regulation	25
12.4 Two Regulatory Worlds: How They Differ	25

12.5 The Geopolitics of Competition.....	26
12.6 A Fragile Multilateralism.....	27
12.7 AI at War	27
13. Human Rights Foundations.....	28
13.1 The Binding Floor.....	28
13.2 The UN Machinery	28
13.3 Scholarship and Its Critics	29
14. The UN System's Evolving Approach to AI	30
14.1 The Usage Track.....	30
14.2 The Governance Track.....	30
14.3 The Political Breakthrough.....	31
14.4 The 2026 Institutions	31
15. The Ethical Tech CoLab: Motivation and Research Philosophy	32
15.1 Motivation	32
15.2 The Portfolio Through Three Lenses.....	32
15.3 Research Philosophy.....	34
16. Conclusion.....	35
Works Cited	36

1. Executive Summary

Humanity has repeatedly transformed itself through technological innovation. Fire, agriculture, metallurgy, writing, navigation, industrialization, electricity, nuclear energy, biotechnology, and digital computing each expanded human capability while simultaneously increasing humanity's capacity to cause harm. Every technological revolution therefore forced societies to develop new ethical frameworks capable of governing newly acquired power. Artificial intelligence represents the latest and perhaps most consequential chapter in this historical process. The ethical questions confronting AI are not unprecedented, because they concern the same enduring challenge that has shaped civilizations for millennia: how should power be exercised responsibly? The issue is power. Ethics is the answer.

This paper deliberately resists the convention, common in the AI ethics literature, of opening with principles such as fairness, transparency, and accountability. Principles detached from their foundations are easily reduced to slogans, and slogans are easily captured by the institutions they are meant to constrain. The paper therefore begins with prior questions. Section 2 frames the oldest of them, humanity's relationship with power, the paradox by which every expansion of capability has expanded the capacity for harm. Sections 3 through 5 ask what ethics is, why human beings created it, and how it evolved from local custom into universal human rights. Section 6 asks what ethical international relations means, drawing on the Carnegie Council's framework of ethical realism and structured ethical deliberation. Section 7 asks what makes technology ethical, and Section 8 arrives, on that foundation, at the question that titles the paper: what is ethical AI? The answer defended throughout is institutional rather than technical. Ethical AI is not a product feature. It is artificial intelligence whose entire lifecycle remains accountable to ethical deliberation, to human rights as a floor, to the do no harm obligation, and to the

inclusive participation of those the technology affects. No system is ethical in itself; systems are made ethical, or fail to be, by the institutions that govern them.

The second half of the paper tests that answer where the stakes are highest. Section 9 traces the rise of the responsible AI movement through its three acts of principles, reckoning, and operationalization. Sections 10 and 11 turn to humanitarian action, tracing the evolution of AI in humanitarian work from the crisis mapping of 2008 to the generative systems of 2026, and showing why humanitarian AI is fundamentally different from commercial AI: consent collapses, experimentation finds the least protected subjects, accountability runs to people with no vote and no purchase, and the cost of failure is measured in lives. Sections 12 through 14 supply the governance context: the geopolitics of AI regulation, the human rights foundations that predate the technology, and the United Nations system's evolving approach, which in 2026 produced the first report of the Independent International Scientific Panel on AI and the first Global Dialogue on AI Governance in Geneva.¹ Section 15 presents the Ethical Tech CoLab, whose portfolio, from civilian evacuation models to a forced labour risk index, puts the paper's argument into practice, and Section 16 concludes.

The paper is written for policymakers, humanitarian practitioners, donors, and researchers rather than AI engineers, in the conviction that the disciplines it draws together, philosophy, ethics, international affairs, humanitarian ethics, digital ethics, and responsible innovation, are not six literatures but one conversation. Societies do not flourish because they possess increasingly powerful tools. They flourish because they develop increasingly trustworthy ways of governing them.

2. Humanity's Relationship with Power

2.1 The Paradox of Capability

The history of civilization is not merely the history of technological innovation. It is the history of humanity learning to govern increasingly powerful forms of power. Every civilization has faced the same paradox. Human beings possess a remarkable capacity to create tools that extend their abilities beyond natural biological limits. Fire transformed survival. Agriculture transformed settlement. Writing transformed memory. Navigation transformed geography. Industrial machinery transformed production. Nuclear physics transformed warfare and medicine. Artificial intelligence now promises to transform decision making itself. Each of these revolutions increased humanity's ability to reshape the world, and every expansion of capability simultaneously increased humanity's capacity to cause harm: new prosperity and discovery on one side, previously unimaginable forms of exploitation, violence, inequality, and domination on the other.

¹ "Global Push for AI Governance amid Warnings of 'Catastrophic Harm,'" *UN News*, 5 July 2026, news.un.org/en/story/2026/07/1167862.

2.2 Power in Social Thought

Power itself is the oldest subject of social thought. Max Weber defined it as the probability that an actor within a social relationship will be in a position to carry out his own will despite resistance, a definition that fits an algorithm allocating aid as readily as a sovereign commanding subjects.² Bertrand Russell went further, arguing that power is the fundamental concept in social science, in the same sense in which energy is the fundamental concept in physics, and that no account of society is complete until it explains how power is distributed and restrained.³ And Lord Acton supplied the warning that every constitutional tradition has since repeated: power tends to corrupt, and absolute power corrupts absolutely.⁴ Read together, these three claims frame the problem this paper addresses. Power is universal, power is the currency of social life, and power unrestrained degrades the one who holds it and endangers everyone subject to it.

2.3 The Recurring Pattern and the New Concentration of Power

History therefore reveals a recurring pattern. Technological innovation alone has never guaranteed human progress. Every major expansion of human capability has required a corresponding expansion of humanity's ethical, legal, and political institutions. Writing required legal systems. Commerce required contracts. Industrialization generated unprecedented prosperity while requiring labor protections, environmental regulation, and public health institutions. Nuclear physics transformed medicine while necessitating international non proliferation regimes. Without such institutions, technological power becomes detached from moral responsibility, and societies confront crises of legitimacy, injustice, and instability that threaten both individuals and the institutions on which collective life depends.

Artificial intelligence is the latest expression of this enduring dynamic. Contemporary discussion often portrays AI as an unprecedented disruption, but its ethical significance lies not in its novelty. It lies in the type of power it creates. AI increasingly participates in decisions concerning healthcare, education, employment, humanitarian assistance, migration, policing, financial services, and national security, and in doing so it redistributes authority between human institutions and computational systems. The fundamental question is therefore not whether AI is intelligent, but whether societies possess institutions capable of governing this new concentration of power responsibly.

This paper argues that ethics has always performed precisely this function. Ethics is not an external constraint imposed on innovation after technological progress has occurred. It is the normative architecture through which societies determine how newly acquired forms of power should be exercised. Every civilization has developed ethical systems because

² Max Weber, *Economy and Society: An Outline of Interpretive Sociology*, edited by Guenther Roth and Claus Wittich, U of California P, 1978, p. 53.

³ Bertrand Russell, *Power: A New Social Analysis*, George Allen and Unwin, 1938, ch. 1.

⁴ John Emerich Edward Dalberg-Acton, letter to Mandell Creighton, 5 Apr. 1887, in *Historical Essays and Studies*, edited by John Neville Figgis and Reginald Vere Laurence, Macmillan, 1907.

every civilization has confronted the same challenge: power without restraint ultimately undermines the conditions that make civilization itself possible. The emergence of artificial intelligence does not alter this lesson. It reinforces it. The ethical questions surrounding AI are neither unprecedented nor exclusively technical. They belong to a much older conversation concerning justice, legitimacy, accountability, and human dignity that has accompanied every major transformation in human history. The sections that follow reconstruct that conversation from its beginning.

3. What Is Ethics?

3.1 Definition and Deliberation

Few concepts have shaped human civilization more profoundly than ethics, yet few are as frequently invoked without careful definition. Governments invoke ethics when justifying military interventions or humanitarian assistance. Corporations establish ethics committees to oversee artificial intelligence and biotechnology. International organizations describe ethical obligations toward refugees, displaced populations, and future generations. Despite its ubiquity, the term is often reduced to a synonym for compliance, legality, or professional conduct. Such simplifications obscure the deeper tradition from which ethics emerged and diminish its significance as a framework for collective decision making.

At its most fundamental level, ethics is the systematic study of how human beings ought to act, what values should guide individual and collective behavior, and how competing moral claims should be evaluated. The Carnegie Council for Ethics in International Affairs puts the point plainly: ethics is about how things ought to be, not simply how things are or are likely to be, and it forces us to look beyond our own immediate interests to consider the interests of others, as well as the values, rules, standards, rights, and obligations that should guide behavior.⁵ Joel Rosenthal, the Council's president, describes ethics as "making the effort to evaluate competing points of view and then truly caring about the impact of the choices you have made," adding that ethics "is not some cure-all for the world's problems, but it is an actual process for finding solutions."⁶ Ethics, on this understanding, is not a fixed doctrine but a disciplined practice of judgment.

3.2 Ethics and Morality

The word itself originates from the ancient Greek term *ethos*, referring to character, custom, and habitual ways of living. Aristotle used the concept to describe the cultivation of virtues that enable individuals to flourish within a political community; for him, ethical excellence was acquired through practice and habituation rather than rule following alone.⁷

⁵ Carnegie Council for Ethics in International Affairs, *Ethics in International Affairs 101*, Carnegie Council, carnegiecouncil.org, pp. 8 to 9.

⁶ Joel Rosenthal, quoted in Carnegie Council, *Ethics in International Affairs 101*, p. 10.

⁷ Aristotle, *Nicomachean Ethics*, translated by Terence Irwin, 2nd ed., Hackett, 1999, bk. II.

Although ethics and morality are frequently used interchangeably, scholarship often distinguishes between them. Morality generally refers to the beliefs, norms, and practices that characterize a particular community. Ethics, by contrast, represents the critical examination of those beliefs. As the anthropologist James Laidlaw observes, the study of ethics concerns not only the moral rules a society enforces but the ways people consciously deliberate about how they ought to live and what kind of persons they aspire to become.⁸ Ethics asks whether prevailing moral norms are justified, whether they produce justice, and whether they should evolve in response to new knowledge or changing circumstances.

This distinction explains why ethical inquiry frequently challenges existing social arrangements rather than reinforcing them. Throughout history, practices once regarded as morally acceptable within particular societies, including slavery, colonial domination, racial segregation, and child labor, were questioned, criticized, and ultimately rejected through ethical reasoning that appealed to principles beyond inherited custom. Ethics is in this sense inherently dynamic. As the Carnegie Council emphasizes, ethics is not static: new relationships between peoples, new scientific findings, and new technologies create new areas for ethical thought that may change past judgments of which we previously felt certain.⁹

3.3 Ethics, Law, and Power

Nor should ethics be confused with law. Legal systems often reflect ethical principles, but legality and morality are neither identical nor perfectly aligned. Apartheid in South Africa and the Nuremberg Laws in Nazi Germany were legally enforceable within their jurisdictions while violating principles now recognized as universal human rights. Conversely, many ethical obligations exceed legal requirements. Humanitarian workers who risk their lives to protect civilians in armed conflict act from commitments that cannot be reduced to statutory duties. Ethical reflection provides the normative foundation upon which legitimate legal systems aspire to rest.

Finally, ethics is not merely individual virtue. Ethical questions emerge wherever human beings exercise power over one another, and they therefore operate simultaneously at several levels: the personal decisions of individuals, the conduct of groups and organizations, and the behavior of states and international institutions.¹⁰ Decisions concerning taxation, migration, public health, humanitarian assistance, military intervention, and artificial intelligence all involve the distribution of benefits, burdens, risks, and opportunities. Technological capability alone offers no guidance as to whether a given application advances justice, respects human dignity, or contributes to the common good. Ethics is the bridge between human capability and human responsibility.

⁸ James Laidlaw, *The Subject of Virtue: An Anthropology of Ethics and Freedom*, Cambridge UP, 2014.

⁹ Carnegie Council, *Ethics in International Affairs* 101, p. 20.

¹⁰ Carnegie Council, *Ethics in International Affairs* 101, pp. 16 to 19.

4. Why Humanity Created Ethics

4.1 Cooperation and Survival

One of the most persistent misconceptions surrounding ethics is the belief that moral philosophy began as an intellectual exercise conducted by ancient philosophers seeking abstract truths. While philosophers transformed ethics into a systematic field of inquiry, ethical behavior itself predates philosophy by thousands of years. Ethics emerged not because human beings suddenly became virtuous, but because societies discovered that power without restraint ultimately destroys the communities upon which power itself depends.

The evolutionary record supports this reading. Human beings survived and flourished not because individuals were stronger than competing species, but because groups developed unusual capacities for communication, collective planning, resource sharing, and long term cooperation. Michael Tomasello's research on the natural history of human morality argues that distinctively human moral psychology, including the sense of fairness, obligation, and shared commitment, arose from the demands of interdependent cooperation: individuals who could not be trusted as partners were left out of the cooperative activities on which survival depended.¹¹ Ethical norms reduced uncertainty in human interaction long before economists developed the vocabulary of transaction costs. They enabled strangers to cooperate, institutions to emerge, and increasingly complex civilizations to develop.

4.2 The Earliest Ethical Systems

The earliest documented normative systems make the connection between ethics and the governance of power explicit. In ancient Egypt, the principle of Maat, embodying truth, justice, balance, and reciprocity, served as the normative foundation upon which political authority derived its legitimacy; rulers were expected not merely to govern effectively but to preserve social equilibrium and protect the vulnerable.¹² In Mesopotamia, the Code of Hammurabi, produced in the eighteenth century BCE, publicly articulated standards of proportionality, responsibility, and protection, and justified them in terms that remain strikingly modern: the laws existed, its prologue declares, so that the strong might not injure the weak.¹³ The transition from unwritten custom to publicly accessible legal standards also anticipated a principle central to contemporary technology governance. Citizens cannot comply with secret laws, and individuals cannot meaningfully contest algorithmic decisions whose logic remains entirely opaque. The modern demand for explainable artificial intelligence reflects a very old ethical commitment to public accountability. In China, Confucius articulated an ethics of cultivated relationships centered

¹¹ Michael Tomasello, *A Natural History of Human Morality*, Harvard UP, 2016.

¹² Maulana Karenga, *Maat, the Moral Ideal in Ancient Egypt: A Study in Classical African Ethics*, Routledge, 2004.

¹³ *The Code of Hammurabi*, translated by L. W. King, 1910, The Avalon Project, Yale Law School, avalon.law.yale.edu/ancient/hamframe.asp.

on ren, or humaneness, in which those who hold power bear heightened responsibilities precisely because their decisions shape the welfare of entire communities.¹⁴

4.3 Ethics as Institutional Infrastructure

Political philosophy generalized these insights. Thomas Hobbes argued that without commonly accepted rules governing behavior, human societies risk descending into a condition of insecurity and perpetual conflict.¹⁵ Later thinkers, including Locke, Rousseau, and Kant, increasingly linked political legitimacy to the protection of rights and reciprocal obligation rather than to coercive power alone. Modern institutional economics reaches a similar conclusion through different methods. Douglass North demonstrated that long term prosperity depends upon institutions capable of generating trust and credible commitments among diverse actors, and Elinor Ostrom showed empirically that communities can govern shared resources sustainably when they develop norms of reciprocity, monitoring, and accountability.¹⁶ Ethics, on this view, functions as institutional infrastructure. It enables complex societies to coordinate collective action without relying exclusively on coercion.

The humanitarian sector illustrates the relationship with particular clarity. Humanitarian organizations frequently operate where formal state institutions have weakened or collapsed. Under such conditions, the principles of humanity, impartiality, neutrality, and independence do not merely guide organizations internally; they generate the trust among affected populations, local authorities, and warring parties on which humanitarian access itself depends. Artificial intelligence introduces a comparable institutional challenge. Refugees cannot audit the machine learning models that help determine assistance eligibility, and displaced families cannot meaningfully contest automated decisions influencing access to shelter or protection. Ethical governance must therefore substitute for informational asymmetries that vulnerable populations cannot realistically overcome themselves. As throughout history, trust cannot be engineered; it must be earned by legitimate institutions operating under shared ethical principles.

5. From Custom to Universal Ethics: The Expansion of Moral Concern

5.1 The Expanding Circle

The history of ethics is, in many respects, the history of an expanding moral community. Early societies focused their strongest obligations on those with whom individuals shared kinship, language, territory, or faith, and outsiders often received fewer protections. This pattern reflects not a lack of ethical sophistication but the practical realities of societies

¹⁴ Confucius, *Analects*, translated by Edward Slingerland, Hackett, 2003.

¹⁵ Thomas Hobbes, *Leviathan*, edited by Richard Tuck, Cambridge UP, 1996, ch. 13.

¹⁶ Douglass C. North, *Institutions, Institutional Change and Economic Performance*, Cambridge UP, 1990; Elinor Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge UP, 1990.

whose institutions rarely extended beyond regional boundaries. As societies became larger and more interconnected, ethical thought expanded with them. The philosopher Peter Singer describes this trajectory as the expanding circle of moral concern, in which the reach of ethical consideration widens from family to tribe, from tribe to nation, and from nation toward humanity as a whole.¹⁷ Kwame Anthony Appiah frames the same process as an ongoing cosmopolitan conversation across difference, in which increased contact among cultures encourages recognition of shared human interests without erasing cultural particularity.¹⁸ Martha Nussbaum argues that ethical reasoning challenges individuals to regard themselves not only as members of particular communities but as citizens of a wider human community whose dignity deserves equal consideration.¹⁹

5.2 The Universal Declaration

The catastrophes of the twentieth century converted this philosophical trajectory into institutional fact. The Holocaust demonstrated that legality and prevailing social morality could become instruments of profound injustice when detached from broader commitments to human dignity. The adoption of the Universal Declaration of Human Rights in 1948 marked the decisive turn: its first article declares that all human beings are born free and equal in dignity and rights, establishing dignity rather than citizenship as the normative foundation of international order.²⁰ For the first time, the international community formally affirmed that certain protections belong to every person by virtue of their humanity alone.

5.3 Universalism Without Uniformity

Universal ethics does not imply ethical uniformity. Contemporary scholarship distinguishes universalism from cultural homogenization: core principles such as protection from torture and recognition of equal dignity apply everywhere, while their practical expression legitimately varies across societies. UNESCO's Recommendation on the Ethics of Artificial Intelligence adopts precisely this position, grounding AI governance in universal human rights commitments while explicitly recognizing the importance of cultural diversity, linguistic plurality, and local participation in technology governance.²¹ The balance matters for humanitarian practice, which requires deep respect for local knowledge alongside an unwavering commitment to internationally recognized rights.

Artificial intelligence intensifies this old tension in a new form. Algorithmic systems are typically developed within particular social, linguistic, and economic contexts before being deployed globally, and they may encode the assumptions of the societies that built them.

¹⁷ Peter Singer, *The Expanding Circle: Ethics, Evolution, and Moral Progress*, Princeton UP, 2011.

¹⁸ Kwame Anthony Appiah, *Cosmopolitanism: Ethics in a World of Strangers*, W. W. Norton, 2006.

¹⁹ Martha C. Nussbaum, "Patriotism and Cosmopolitanism," *Boston Review*, vol. 19, no. 5, 1994.

²⁰ United Nations General Assembly, *Universal Declaration of Human Rights*, Resolution 217 A (III), 10 Dec. 1948, art. 1.

²¹ UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, UNESCO, 23 Nov. 2021, unesco.org/en/artificial-intelligence/recommendation-ethics.

Research on algorithmic fairness has demonstrated that models trained primarily on data from wealthy, highly connected populations frequently perform worse for marginalized communities and underrepresented languages.²² These disparities are not merely technical shortcomings. They raise ethical questions about whose experiences are represented, whose knowledge is valued, and whose rights receive effective protection. The evolution from local morality toward universal ethics thus provides the essential test for AI governance: a technically sophisticated system that systematically disadvantages displaced populations or excludes minority languages may satisfy engineering benchmarks while failing the standards on which humanitarian action itself is founded.

6. What Is Ethical International Relations? Realism, Idealism, and Ethical Realism

6.1 Realism and Idealism

International affairs pose the hardest case for ethics, because they are the arena in which power is least constrained. Two extreme views have long competed. Realism, in its strongest version, holds that states pursue power and narrowly defined self interest, and that ethics has little to contribute; Henry Kissinger captured its spirit in the warning that “a country that demands moral perfection in its foreign policy will achieve neither perfection nor security.”²³ Idealism holds that what matters is adherence to principles and international law, in the lineage of Kant’s proposal for a federation of states in Perpetual Peace, which inspired the League of Nations and, later, the United Nations.²⁴ Each view captures something real, and each is insufficient on its own. States plainly do pursue interests; norms plainly do change behavior. As Rosenthal observes, realism alone cannot account for the enormous shifts in expected and required behavior over the past century, including prohibitions on child labor, expectations of equal treatment for women, and duties toward the natural world.²⁵

6.2 Ethical Realism

The Carnegie Council’s answer, and the orientation this paper adopts, is ethical realism: a view of morality in global politics that takes seriously both the role of ethics and the nature of power and its relationship to national interests.²⁶ Hans Morgenthau, the founder of modern realism, himself framed the choice this way: not between moral principles and the

²² Joy Buolamwini and Timnit Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” *Proceedings of Machine Learning Research*, vol. 81, 2018, pp. 77 to 91.

²³ Henry Kissinger, *Foreign Affairs*, June 1994, quoted in Carnegie Council, *Ethics in International Affairs* 101, p. 24.

²⁴ Immanuel Kant, “Toward Perpetual Peace” (1795), in *Practical Philosophy*, translated and edited by Mary J. Gregor, Cambridge UP, 1996; see Carnegie Council, *Ethics in International Affairs* 101, pp. 26 to 27.

²⁵ Joel Rosenthal, quoted in Carnegie Council, *Ethics in International Affairs* 101, p. 25.

²⁶ Carnegie Council, *Ethics in International Affairs* 101, pp. 5 and 29.

national interest devoid of moral dignity, “but between one set of moral principles divorced from political reality and another set of moral principles derived from political reality.”²⁷ Ethical realism refuses both the cynicism that treats ethics as decoration and the utopianism that ignores power. It asks what the most ethical policy is that is also possible.

6.3 Ethical Deliberation as Method

Ethical realism operates through a structured process of ethical deliberation. In the Carnegie Council’s formulation, deliberation involves identifying goals or outcomes, finding means to satisfy them, and assessing and revising both goals and means against standards, values, principles, duties, and rights; the cycle runs from identifying an ethical problem through deliberation to choice and action, and then to reflection and the taking of responsibility.²⁸ Two features of this process deserve emphasis for what follows. First, deliberation is pluralistic. The main Western frameworks, consequentialism, virtue ethics, and rights or duty based ethics, each illuminate a different aspect of a decision, and solving real problems may require drawing on each without endorsing any single framework as uniquely correct.²⁹ Second, deliberation is inclusive. It is best done with others, especially those most affected by an action or policy, because on our own we cannot see the full picture and we tend to be biased toward decisions that benefit us.³⁰ These two commitments, pluralism and inclusion, recur throughout this paper. They are the methodological core of what the Ethical Tech CoLab means by ethical technology, and they are the standard against which the AI governance frameworks examined below should be judged.

7. What Is Ethical Tech?

7.1 Artifacts Have Politics

Before the question of ethical AI can be answered, a broader one deserves a direct answer, because the CoLab carries it in its name: what makes technology ethical? The scholarly tradition here is older than the AI debate and clearer than is often assumed. Its founding insight is that technologies are not neutral instruments whose moral character depends entirely on their users. Langdon Winner’s classic essay asked whether artifacts have politics and answered that they do: the design of a bridge, a machine, or a system can embody and enforce social arrangements, settling questions of power before any user makes a choice.³¹ James Moor, founding the field of computer ethics, observed that

²⁷ Hans J. Morgenthau, *Politics Among Nations: The Struggle for Power and Peace*, Alfred A. Knopf, 1948, quoted in Carnegie Council, *Ethics in International Affairs* 101, p. 28.

²⁸ Carnegie Council, *Ethics in International Affairs* 101, pp. 12 to 13.

²⁹ Carnegie Council, *Ethics in International Affairs* 101, pp. 14 to 15.

³⁰ Carnegie Council, *Ethics in International Affairs* 101, p. 12.

³¹ Langdon Winner, “Do Artifacts Have Politics?” *Daedalus*, vol. 109, no. 1, 1980, pp. 121 to 136.

computing creates policy vacuums, situations in which technology confers new capabilities faster than societies can form rules for their use, and that filling those vacuums is a conceptual and ethical task, not a technical one.³² Every stage of technological development reflects judgments about whose interests deserve protection, which values are optimized, and which trade offs are acceptable. Designers decide which objectives a system pursues; organizations set acceptable levels of risk; investors shape research priorities; data practices shape behavior. Technology is, in this sense, frozen judgment.

7.2 Value Sensitive Design and Responsible Innovation

Ethical technology, then, is technology whose embedded judgments have been made deliberately, inclusively, and accountably. Two research programs give that definition operational content. Value sensitive design, developed by Batya Friedman and colleagues, provides a method for identifying the stakeholders of a system, direct and indirect, surfacing the values at stake, and building those values into the design rather than auditing for them afterward.³³ The responsible research and innovation literature, associated with Jack Stilgoe, Richard Owen, and Phillip Macnaghten, adds four commitments that any innovating institution can be held to: anticipation of plausible harms, reflexivity about the innovator's own assumptions, inclusion of affected publics in shaping the trajectory, and responsiveness, meaning the demonstrated capacity to change course when evidence or deliberation demands it.³⁴ These frameworks predate the current AI moment, and that is their value. They establish that the ethics of a technology is decided in its governance, upstream in design and downstream in oversight, and not in its marketing.

7.3 Ethical Tech as Practice

This is the sense in which the CoLab uses the term ethical tech. It does not name a category of inherently benign products. It names a practice: choosing problems by the harm they address rather than the market they open, designing with the affected rather than for them, embedding safeguards as constraints rather than aspirations, and accepting accountability structures with the power to stop a build. The definition of ethical AI developed in the next section is a special case of this practice, applied to the most consequential technology of the present decade.

³² James H. Moor, "What Is Computer Ethics?" *Metaphilosophy*, vol. 16, no. 4, 1985, pp. 266 to 275.

³³ Batya Friedman and David G. Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, MIT Press, 2019.

³⁴ Jack Stilgoe, Richard Owen, and Phillip Macnaghten, "Developing a Framework for Responsible Innovation," *Research Policy*, vol. 42, no. 9, 2013, pp. 1568 to 1580.

8. What Is Ethical AI?

8.1 A Definition

With these foundations in place, the question that titles this paper can be answered without slogans. Ethical AI is not a product feature, a compliance checkbox, or a marketing claim. It is artificial intelligence whose entire lifecycle, from problem selection and data collection through design, deployment, and retirement, remains accountable to ethical deliberation: to human rights as a floor, to the humanitarian principle of doing no harm, and to the inclusive participation of those the technology affects. On this definition, no system is ethical in itself. Systems are made ethical, or fail to be, by the institutions that govern them.

8.2 From Principles to Governance

The field's own history supports this institutional definition. When Anna Jobin, Marcello Ienca, and Effy Vayena mapped eighty four AI ethics guidelines published worldwide, they found an apparent global convergence around five principles, transparency, justice and fairness, non maleficence, responsibility, and privacy, but substantive divergence over how those principles were interpreted, why they were deemed important, to whom they applied, and how they should be implemented.³⁵ They also documented the near absence of the Global South from the production of the guidelines themselves, a finding that should unsettle any claim that the principles are universal in origin rather than merely universal in ambition. Luciano Floridi and Josh Cowls, synthesizing the same landscape, proposed that the recurring principles map closely onto the classical principles of bioethics, beneficence, non maleficence, autonomy, and justice, with one addition specific to AI: explicability, the requirement that systems be intelligible and that someone be answerable for them.³⁶

The limits of principles became apparent almost immediately. Brent Mittelstadt's influential critique observed that AI development lacks precisely what makes principlism work in medicine: common aims and fiduciary duties, professional history and norms, proven methods to translate principles into practice, and robust legal and professional accountability mechanisms.³⁷ A principle that no profession is bound to, no method operationalizes, and no institution enforces is an aspiration, not a governance instrument. This is why the definition offered above locates ethical AI in governance rather than in intention. It is also why the intergovernmental instruments matter. The OECD AI Principles, adopted in 2019 as the first intergovernmental AI standard and amended in 2024 to address general purpose AI, commit adherents to AI actors who respect the rule of law,

³⁵ Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence*, vol. 1, 2019, pp. 389 to 399.

³⁶ Luciano Floridi and Josh Cowls, "A Unified Framework of Five Principles for AI in Society," *Harvard Data Science Review*, vol. 1, no. 1, 2019.

³⁷ Brent Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI," *Nature Machine Intelligence*, vol. 1, 2019, pp. 501 to 507.

human rights, and democratic values throughout the AI system lifecycle.³⁸ The IEEE's Ethically Aligned Design brought the engineering profession into the conversation and seeded a family of technical standards, demonstrating that ethical requirements can be translated into specification.³⁹ UNESCO's Recommendation on the Ethics of Artificial Intelligence, adopted by all of the organization's member states in 2021, grounds AI governance in human rights, dignity, environmental sustainability, and cultural diversity, and pairs values with implementation tools.⁴⁰ These instruments do not invent new moral values for the digital age. They translate enduring ethical principles into practical guidance, which is exactly what ethics has done at every previous technological transition.

8.3 Accuracy Is Not Ethics

One further clarification completes the definition. Ethical AI is frequently conflated with accurate AI, as though sufficient technical performance would resolve the ethical questions. It would not. A perfectly accurate system pursuing an unethical objective remains unethical, and history repeatedly demonstrates that efficiency without legitimacy may strengthen injustice rather than reduce it. The question ethics asks of an AI system is the question Carnegie Council's deliberative framework asks of any policy: not only whether the means work, but whether the ends are justified, and whether the relationship between ends and means respects the dignity of those affected.⁴¹

9. The Rise of Responsible AI

9.1 The Age of Principles

The movement now called responsible AI emerged in three acts, and each act refined the answer to what ethical AI requires. The first act was the age of principles. Between roughly 2016 and 2019, ethics codes issued from every quarter: the Asilomar AI Principles from the research community, which declared that the goal of AI research should be to create not undirected intelligence but beneficial intelligence; the Montreal Declaration, distinctive for its methodology of citizen deliberation; the IEEE's engineering framework; and the OECD Principles discussed above.⁴² The proliferation itself was evidence of sincerity, but also of fragmentation, as the Jobin study showed.

³⁸ Organisation for Economic Co-operation and Development, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449, OECD, 2019, amended 2024.

³⁹ IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, 1st ed., IEEE, 2019.

⁴⁰ UNESCO, *Recommendation on the Ethics of Artificial Intelligence*.

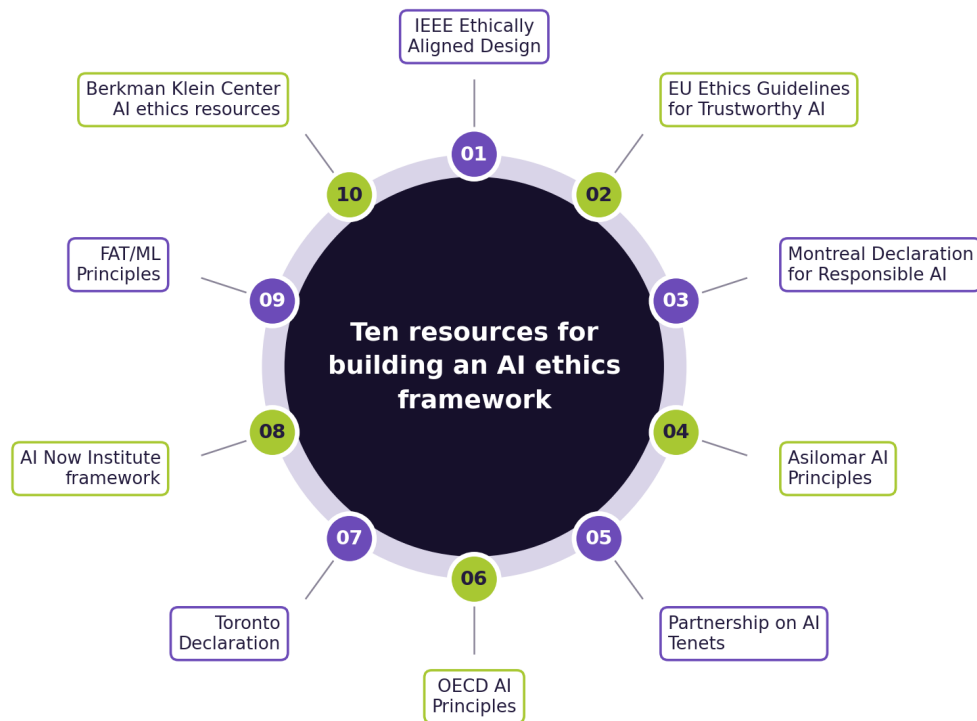
⁴¹ Carnegie Council, *Ethics in International Affairs 101*, pp. 12 to 15.

⁴² Future of Life Institute, "Asilomar AI Principles," Future of Life Institute, 2017, futureoflife.org/open-letter/ai-principles; Université de Montréal, *Montreal Declaration for a Responsible Development of Artificial Intelligence*, 2018, montrealdeclaration-responsibleai.com.

9.2 Ten Landmark Frameworks

The first act left behind a canon. Ten resources, compiled usefully in a practitioner survey by Emily Johnson for SecureITWorld, remain the standard references for any institution building an AI ethics framework, and each contributes something distinct.⁴³ Figure 4 presents them together; the paragraphs that follow explain each in turn.

Figure 4. Ten landmark resources for building an AI ethics framework



Adapted from Emily Johnson, "AI Ethics Frameworks: 10 Essential Resources," SecureITWorld, 12 March 2025.

Figure 4. Ten landmark resources for building an AI ethics framework.

The IEEE's Ethically Aligned Design is the engineering profession's contribution: an extensive handbook on prioritizing human wellbeing in autonomous and intelligent systems that seeded the IEEE P7000 series of technical standards, demonstrating that ethical requirements can be translated into specification.⁴⁴ The European Union's Ethics Guidelines for Trustworthy AI, prepared by the Commission's High Level Expert Group in

⁴³ Emily Johnson, "AI Ethics Frameworks: 10 Essential Resources to Build an Ethical AI Framework," SecureITWorld, 12 Mar. 2025, secureitworld.com/blog/ai-ethics-frameworks-10-essential-resources-to-build-an-ethical-ai-framework.

⁴⁴ IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, 1st ed., IEEE, 2019.

2019, defined trustworthy AI as lawful, ethical, and robust, and set out seven requirements, from human agency and oversight to accountability, that later shaped the EU AI Act.⁴⁵ The Montreal Declaration for the Responsible Development of Artificial Intelligence is distinctive for its method as much as its content: its ten principles, including wellbeing, privacy, solidarity, and sustainability, emerged from an extended public deliberation with citizens, a model of the inclusive participation this paper has argued ethics requires.⁴⁶

The Asilomar AI Principles, adopted at the 2017 Future of Life Institute conference, articulated twenty three principles spanning research culture, ethics and values, and longer term issues, and framed the goal of the field as beneficial rather than merely capable intelligence.⁴⁷ The Partnership on AI's founding Tenets committed an unusual coalition, major technology companies alongside civil society organizations, to safety, fairness, transparency, and accountability, and to the proposition that AI should benefit and empower as many people as possible.⁴⁸ The OECD AI Principles, examined earlier in this paper, remain the first intergovernmental AI standard and the reference point for national policy; their emphasis on inclusive growth, human centred values, and transparency makes them the natural starting point for governments.⁴⁹

The Toronto Declaration, launched in 2018 by Amnesty International and Access Now, is the human rights community's entry: it applies the existing law of equality and non discrimination to machine learning systems, insisting on effective remedy for those harmed, and anticipates the rights based argument developed in Section 13 of this paper.⁵⁰ The AI Now Institute, founded at New York University, contributed a research program rather than a code: its annual reports pressed the field on accountability, labor, and bias as they operate in deployed systems, and its recommendation of algorithmic impact assessments for public agencies has been widely adopted.⁵¹ The FAT/ML community's Principles for Accountable Algorithms, with their accompanying social impact statement, gave individual developers a concrete instrument: five commitments, responsibility, explainability, accuracy, auditability, and fairness, addressed to the people who actually

⁴⁵ High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, 8 Apr. 2019.

⁴⁶ Future of Life Institute, "Asilomar AI Principles," Future of Life Institute, 2017, futureoflife.org/open-letter/ai-principles; Université de Montréal, *Montreal Declaration for a Responsible Development of Artificial Intelligence*, 2018, montrealdeclaration-responsibleai.com.

⁴⁷ Future of Life Institute, "Asilomar AI Principles," Future of Life Institute, 2017, futureoflife.org/open-letter/ai-principles; Université de Montréal, *Montreal Declaration for a Responsible Development of Artificial Intelligence*, 2018, montrealdeclaration-responsibleai.com.

⁴⁸ Partnership on AI, "Our Tenets," Partnership on AI, 2016, partnershiponai.org.

⁴⁹ Organisation for Economic Co-operation and Development, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449, OECD, 2019, amended 2024.

⁵⁰ Amnesty International and Access Now, *The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems*, 16 May 2018.

⁵¹ Meredith Whittaker et al., *AI Now Report 2018*, AI Now Institute, New York University, Dec. 2018.

build the systems.⁵² And the Berkman Klein Center at Harvard contributed the map of the whole terrain: its Principled Artificial Intelligence report analyzed thirty six prominent AI principles documents and identified eight recurring themes, providing the empirical evidence for both the convergence and the gaps this section has described.⁵³

Read as a set, the ten resources trace the same arc as this paper: from professional self regulation through public deliberation and intergovernmental standards to human rights law, and from principles addressed to institutions toward instruments usable by the practitioners who build and the communities who bear the consequences.

9.3 The Reckoning

The second act was the reckoning, and its evidence was empirical. Joy Buolamwini and Timnit Gebru's Gender Shades study demonstrated that commercial facial analysis systems misclassified darker skinned women at error rates approaching 34.7 percent while the maximum error rate for lighter skinned men was 0.8 percent, a single result that legitimized the independent algorithmic audit as a discipline.⁵⁴ Safiya Umoja Noble showed how search algorithms reinforce racism through what she termed technological redlining.⁵⁵ Ruha Benjamin diagnosed the New Jim Code, discriminatory design that encodes inequity while appearing neutral or even benevolent, a warning with particular force for the aid sector, where technological fixes marketed as humanitarian can deepen the harms they claim to solve.⁵⁶ Kate Crawford's Atlas of AI reframed the technology as an extractive industry whose costs in minerals, labor, and data fall far from its beneficiaries.⁵⁷ And Elettra Bietti named the political economy of the moment, ethics washing, the instrumental use of ethics language by industry to forestall binding regulation, while cautioning equally against dismissing ethics wholesale.⁵⁸

9.4 Operationalization and the Safety Turn

The third act was operationalization. The critique produced tools: model cards for standardized reporting of model performance across demographic and environmental conditions; datasheets for datasets, documenting motivation, composition, and recommended use; and end to end internal audit frameworks adapted from aerospace and

⁵² Nicholas Diakopoulos et al., "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms," Fairness, Accountability, and Transparency in Machine Learning (FAT/ML), fatml.org/resources/principles-for-accountable-algorithms.

⁵³ 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*, Berkman Klein Center Research Publication no. 2020-1, Harvard University, 2020.

⁵⁴ Buolamwini and Gebru, "Gender Shades," pp. 77 to 91.

⁵⁵ Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York UP, 2018.

⁵⁶ Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code*, Polity, 2019.

⁵⁷ Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale UP, 2021.

⁵⁸ Elettra Bietti, "From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM, 2020, pp. 210 to 219.

finance.⁵⁹ It then produced institutions. The United States National Institute of Standards and Technology released its AI Risk Management Framework in January 2023, organizing practice into four functions, govern, map, measure, and manage.⁶⁰ ISO and IEC published ISO/IEC 42001 in December 2023, the first certifiable AI management system standard, completing responsible AI's journey from manifesto to auditable compliance.⁶¹ From late 2023 the center of gravity shifted again, from fairness in deployed systems to safety in frontier models. The Bletchley Declaration, signed by twenty eight countries and the European Union including both the United States and China, acknowledged the potential for serious, even catastrophic, harm stemming from the most significant capabilities of frontier models, and launched a summit series and a network of national AI safety institutes.⁶² The International AI Safety Report, prepared by an expert panel chaired by Yoshua Bengio with contributors from more than thirty countries, became the scientific reference document of this safety era.⁶³

The humanitarian reading of this decade is the one on which its sharpest critics converge: principles without accountability structures protect institutions, not people. That lesson, which the humanitarian sector learned independently through its own biometrics controversies, is now constitutive of how humanitarian actors approach AI, and it is the hinge on which this paper turns from the general question of ethical AI to the specific domain of humanitarian action.

10. The Historical Evolution of AI in Humanitarian Work

10.1 Crowdsourcing Origins

The story of AI in humanitarian action does not begin with algorithms but with people. In the aftermath of Kenya's disputed 2008 election, a small group of technologists built Ushahidi, a platform for crowdsourcing reports of violence by text message and web. Two years later, when a magnitude 7.0 earthquake devastated Haiti, volunteers around the world adapted the same model to map urgent needs in near real time, while a parallel effort, Mission 4636, mobilized thousands of Kreyol speaking diaspora volunteers to

⁵⁹ Margaret Mitchell et al., "Model Cards for Model Reporting," *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2019, pp. 220 to 229; Timnit Gebru et al., "Datasheets for Datasets," *Communications of the ACM*, vol. 64, no. 12, 2021, pp. 86 to 92; Inioluwa Deborah Raji et al., "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM, 2020, pp. 33 to 44.

⁶⁰ National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, U.S. Department of Commerce, Jan. 2023.

⁶¹ International Organization for Standardization and International Electrotechnical Commission, *ISO/IEC 42001:2023, Information Technology, Artificial Intelligence, Management System*, ISO, Dec. 2023.

⁶² "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1 and 2 November 2023," GOV.UK, 1 Nov. 2023.

⁶³ Yoshua Bengio et al., *International AI Safety Report 2026*, AI Security Institute and international expert panel, Feb. 2026, internationalaisafetyreport.org.

translate and triage tens of thousands of text messages from survivors.⁶⁴ Patrick Meier, one of the movement's founders, chronicled this era of digital humanitarians, in which affected populations became information producers rather than passive recipients of aid.⁶⁵ Scholars soon consolidated the field under the banner of crisis informatics, treating social media and mobile data as new instruments for understanding disaster while warning from the outset about the limits of interpretation.⁶⁶

Crucially, these early systems were hybrids of human and machine intelligence. Mission 4636's coordinator, Robert Munro, later showed how the human computation of volunteer translators became the training substrate for early natural language processing tools in crisis response.⁶⁷ The artificial intelligence of the humanitarian sector was built on the labor and knowledge of affected communities and volunteers from the start, a genealogy with ethical implications the sector is still working through.

10.2 Institutionalization and Prediction

Institutionalization followed. In 2009 the UN Secretary General launched Global Pulse, an innovation initiative premised on the idea that the digital traces of mobile phone use and online activity could provide real time insight into human wellbeing.⁶⁸ By the mid 2010s, the big data era had given way to operational machine learning. The World Food Programme's HungerMap LIVE began nowcasting food insecurity across dozens of countries from mobile surveys, satellite imagery, and market data.⁶⁹ The World Bank's Famine Action Mechanism linked machine learning famine prediction to pre arranged financing.⁷⁰ OCHA's Centre for Humanitarian Data built a peer review framework for predictive analytics that allowed the UN's Central Emergency Response Fund to release money before forecasted floods struck Bangladesh in 2020.⁷¹

10.3 Identification and Biometrics

Alongside prediction came identification. UNHCR had experimented with iris scan registration of Afghan returnees as early as 2002, and by the late 2010s biometric

⁶⁴ Jessica Heinzelman and Carol Waters, *Crowdsourcing Crisis Information in Disaster-Affected Haiti*, Special Report 252, United States Institute of Peace, 2010.

⁶⁵ Patrick Meier, *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response*, CRC Press, 2015.

⁶⁶ Leysia Palen and Kenneth M. Anderson, "Crisis Informatics: New Data for Extraordinary Times," *Science*, vol. 353, no. 6296, 2016, pp. 224 to 225.

⁶⁷ Robert Munro, "Crowdsourcing and the Crisis-Affected Community: Lessons Learned and Looking Forward from Mission 4636," *Information Retrieval*, vol. 16, no. 2, 2013, pp. 210 to 266.

⁶⁸ UN Global Pulse, *Big Data for Development: Challenges and Opportunities*, United Nations, May 2012.

⁶⁹ World Food Programme, "WFP Releases HungerMap LIVE," WFP, 2019, wfp.org/news/wfp-releases-hungermap-live-modernized-intelligence-platform-turns-data-global-hunger-early.

⁷⁰ World Bank, "Famine Action Mechanism (FAM)," World Bank, 2018, worldbank.org/en/programs/famine-early-action-mechanism.

⁷¹ Centre for Humanitarian Data, *Peer Review Framework for Predictive Analytics in Humanitarian Response*, UN OCHA, Mar. 2020.

registration had become standard infrastructure across UNHCR and WFP operations.⁷² In Jordan's Azraq and Zaatari camps, WFP's Building Blocks system coupled iris scan authentication to a permissioned blockchain, allowing refugees to buy groceries with a glance.⁷³ These systems, celebrated by their builders as efficient and fraud resistant, would become the focal point of the sector's fiercest ethical controversies.

10.4 The Generative Era

The generative era arrived abruptly. Sarah Spencer's matched pair of Humanitarian Practice Network papers, published in 2021 and 2024, bookend the arrival of large language models and document a sector scrambling to absorb them into casework, information services, and back office functions.⁷⁴ The most rigorous published evaluation to date, the Signpost AI pilot led by the International Rescue Committee with partners across four countries, tested a large language model information assistant for refugees and migrants: response quality improved markedly over the pilot, and moderator handling time for complex queries fell from as much as twenty five minutes to under five, but roughly fifteen percent of responses could not be rated safe, leading the evaluators to conclude that the tool is not safe without a human in the loop.⁷⁵ At planetary scale, frontier laboratory AI entered humanitarian early warning: Google's flood forecasting model, published in *Nature* in 2024, extended seven day flood warnings to ungauged river basins in more than eighty countries, and promptly generated its own scholarly debate about when AI models are truly ready for deployment.⁷⁶

Three features of this history deserve emphasis. First, the arc runs from participation toward prediction and automation, a shift that progressively distances affected people from the systems that make decisions about them. Second, each technological wave arrived with private sector partners, embedding commercial actors ever deeper in the humanitarian data ecosystem. Third, governance consistently lagged deployment: the data responsibility guidelines and AI policies examined below were built after the systems they now seek to govern were already in the field.

⁷² The Engine Room and Oxfam, *Biometrics in the Humanitarian Sector*, The Engine Room, 2018, updated 2023.

⁷³ Russ Juskalian, "Inside the Jordan Refugee Camp That Runs on Blockchain," *MIT Technology Review*, 12 Apr. 2018.

⁷⁴ Sarah W. Spencer, *Humanitarian AI: The Hype, the Hope and the Future*, Network Paper 85, Humanitarian Practice Network, ODI, Nov. 2021; Sarah W. Spencer, *Humanitarian AI Revisited: Seizing the Potential and Sidestepping the Pitfalls*, Network Paper 89, Humanitarian Practice Network, ODI, May 2024.

⁷⁵ Signpost AI, "Pilot Report: Signpost AI Information Assistant," International Rescue Committee, 27 June 2025, signpostai.org/research/pilot-report-signpost-ai-information-assistant.

⁷⁶ Grey Nearing et al., "Global Prediction of Extreme Floods in Ungauged Watersheds," *Nature*, vol. 627, 2024, pp. 559 to 563.

11. Why Humanitarian AI Is Fundamentally Different from Commercial AI

11.1 Principles Under Pressure

It is tempting to treat humanitarian AI as commercial AI with a nobler mission statement. That view is mistaken, and the difference is structural rather than rhetorical. It begins with the humanitarian principles of humanity, neutrality, impartiality, and independence, which are not aspirational values but operational commitments on which access to conflict zones, the trust of affected communities, and the safety of aid workers depend. As the ICRC's Pierrick Devidal argues, the sector's reflex that if technology can help, it should be used risks quietly subordinating those principles to innovation logics imported from the commercial world.⁷⁷

11.2 Consent and Power

The deepest difference concerns consent and power. Commercial AI at least nominally rests on user agreement: one clicks accept, however illusory the choice. In humanitarian settings even that fiction collapses. A person who must submit to iris scanning to feed her children has not consented in any meaningful sense; she has complied. Mark Latonero named this dynamic surveillance humanitarianism, defined as “the enormous data collection systems deployed by aid organizations that inadvertently increase the vulnerability of people in urgent need,” writing in the wake of the 2019 standoff in Yemen, where WFP suspended aid amid a dispute over biometric registration.⁷⁸ The ICRC's own Handbook on Data Protection in Humanitarian Action concedes that consent is frequently an unusable legal basis in displacement contexts, forcing agencies onto alternative grounds such as vital interest, a candid admission that the commercial consent model simply does not transfer.⁷⁹

11.3 Humanitarian Experimentation

A second structural difference is the asymmetry of experimentation. Kristin Bergtora Sandvik, Katja Lindskov Jacobsen, and Sean Martin McDonald have developed a taxonomy of humanitarian experimentation, showing how untested technologies are routinely piloted on crisis affected populations without the ethics infrastructure that governs experimentation elsewhere: “There is a stark ethical and practical difference between managing risk and introducing it.”⁸⁰ Jacobsen's study of UNHCR's early biometric registration of Afghan refugees documents precisely such an episode, an immature

⁷⁷ Pierrick Devidal, “Lost in Digital Translation? The Humanitarian Principles in the Digital Age,” *International Review of the Red Cross*, vol. 106, no. 925, 2024.

⁷⁸ Mark Latonero, “Stop Surveillance Humanitarianism,” *The New York Times*, 11 July 2019.

⁷⁹ Christopher Kuner and Massimo Marelli, editors, *Handbook on Data Protection in Humanitarian Action*, 2nd ed., ICRC and Brussels Privacy Hub, 2020.

⁸⁰ Kristin Bergtora Sandvik, Katja Lindskov Jacobsen, and Sean Martin McDonald, “Do No Harm: A Taxonomy of the Challenges of Humanitarian Experimentation,” *International Review of the Red Cross*, vol. 99, no. 904, 2017.

technology deployed first on people with the least power to refuse it.⁸¹ Mirca Madianou presses the critique further with the concept of technocolonialism, arguing that digital innovation and data practices in aid rework colonial relationships of extraction, with affected people's data generating value for agencies, states, and technology firms while the risks remain with the data subjects.⁸²

11.4 Accountability and the Cost of Failure

Third, the accountability geometry is inverted. A commercial firm answers, however imperfectly, to customers, shareholders, and regulators in the jurisdictions where it operates. A humanitarian organization's primary obligation runs to affected populations who are neither customers nor voters, who often cannot see the algorithm that ranked their vulnerability, and who typically have no avenue of redress. This is why the sector's own standards, from the OCHA Data Responsibility Guidelines to the Inter Agency Standing Committee's system wide operational guidance, center the do no harm obligation rather than product liability.⁸³ Scholars writing from within the UN system, including Michael Pizzi, Mila Romanoff, and Tim Engelhardt, draw the consequence: human rights law, not voluntary ethics, must govern humanitarian AI across its lifecycle, because bias, opacity, and accountability gaps strike vulnerable populations hardest.⁸⁴ Ana Beduschi's widely cited assessment reaches the operational conclusion that humanitarian AI requires impact assessments driven by the do no harm principle and data protection by design as preconditions of deployment rather than afterthoughts.⁸⁵

Finally, the cost of failure is categorically different. A mistargeted advertisement wastes attention; a mistargeted famine forecast, a false fraud flag on a food ration, or a leaked refugee database can cost lives. Giulio Coppi and colleagues argue that explainability is therefore not optional in humanitarian AI: accountability to affected populations is a sector wide obligation, and an unexplainable system cannot discharge it.⁸⁶ The ICRC's 2024 institutional AI policy codifies the resulting posture: a precautionary approach that permits deployment only with evidence of positive impact, mandates human oversight, preserves non digital alternatives, and commits to transparency about when, why, and how AI is

⁸¹ Katja Lindskov Jacobsen, "Experimentation in Humanitarian Locations: UNHCR and Biometric Registration of Afghan Refugees," *Security Dialogue*, vol. 46, no. 2, 2015, pp. 144 to 164.

⁸² Mirca Madianou, "Technocolonialism: Digital Innovation and Data Practices in the Humanitarian Response to Refugee Crises," *Social Media + Society*, vol. 5, no. 3, 2019, pp. 1 to 13; Mirca Madianou, *Technocolonialism: When Technology for Good Is Harmful*, Polity, 2024.

⁸³ UN OCHA, *OCHA Data Responsibility Guidelines*, Centre for Humanitarian Data, Oct. 2021; Inter-Agency Standing Committee, *IASC Operational Guidance on Data Responsibility in Humanitarian Action*, IASC, 2023.

⁸⁴ Michael Pizzi, Mila Romanoff, and Tim Engelhardt, "AI for Humanitarian Action: Human Rights and Ethics," *International Review of the Red Cross*, vol. 102, no. 913, 2020, pp. 145 to 180.

⁸⁵ Ana Beduschi, "Harnessing the Potential of Artificial Intelligence for Humanitarian Action: Opportunities and Risks," *International Review of the Red Cross*, vol. 104, no. 919, 2022, pp. 1149 to 1169.

⁸⁶ Giulio Coppi, Rebeca Moreno Jimenez, and Sofia Kyriazi, "Explicability of Humanitarian AI: A Matter of Principles," *Journal of International Humanitarian Action*, vol. 6, art. 19, 2021.

used.⁸⁷ No commercial AI company operates under comparable self imposed constraints, because no commercial AI company needs to. That is the difference.

12. The International Affairs Context

12.1 Three Digital Empires

Humanitarian AI does not unfold in a geopolitical vacuum. The technologies that aid agencies adopt, the rules that constrain them, and the funding that sustains them are all shaped by a contest among what Anu Bradford calls the three digital empires: the American market driven model, the Chinese state driven model, and the European rights driven model, each competing to project its regulatory vision globally.⁸⁸

The European Union struck first and most comprehensively. The EU Artificial Intelligence Act, in force since August 2024, establishes the world's first horizontal, risk based AI regime, prohibiting practices such as social scoring outright and designating as high risk a list of uses that includes migration and asylum management, a category of direct humanitarian relevance.⁸⁹ Scholars have nonetheless questioned whether the Act's product safety architecture, which delegates fundamental rights questions to technical standards bodies, can deliver the protection it promises.⁹⁰ The Council of Europe's Framework Convention on Artificial Intelligence, the first legally binding international AI treaty, opened for signature in September 2024 and joined by non European states including the United States, extends the rights based model into international law, though critics note its broad national security carve outs.⁹¹

12.2 The EU AI Act

The European Union's Artificial Intelligence Act deserves closer examination, because it is the world's first comprehensive AI law and the reference point against which every other approach is measured. The Act regulates by risk tier. At the top sit prohibited practices, including social scoring, untargeted scraping of facial images, and emotion recognition in workplaces and schools; below them, high risk systems, a category defined in Annex III that includes AI used in employment, education, essential services, law enforcement, and migration, asylum, and border management, subject to conformity assessment, risk

⁸⁷ International Committee of the Red Cross, *Building a Responsible Humanitarian Approach: The ICRC's Policy on Artificial Intelligence*, ICRC, Nov. 2024.

⁸⁸ Anu Bradford, *Digital Empires: The Global Battle to Regulate Technology*, Oxford UP, 2023.

⁸⁹ European Parliament and Council of the European Union, *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*, Official Journal of the European Union, 12 July 2024, art. 5 and annex III.

⁹⁰ Michael Veale and Frederik Zuiderveen Borgesius, "Demystifying the Draft EU Artificial Intelligence Act," *Computer Law Review International*, vol. 22, no. 4, 2021, pp. 97 to 112.

⁹¹ Council of Europe, *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, Council of Europe Treaty Series no. 225, 5 Sept. 2024.

management, data governance, human oversight, and registration before they reach the market; then limited risk systems carrying transparency duties; and finally minimal risk systems left unregulated. Penalties reach seven percent of global turnover or thirty five million euros for prohibited practices.⁹² The Act entered into force on 1 August 2024 and has been applying in stages: the prohibitions and AI literacy duties since 2 February 2025, and the obligations for general purpose AI models, supported by a Code of Practice, since 2 August 2025.⁹³

The year 2026 brought a significant recalibration. Under sustained pressure over European competitiveness, the Commission's Digital Omnibus package, provisionally agreed on 6 May 2026, postponed the compliance deadlines for high risk systems, moving stand alone Annex III systems from 2 August 2026 to 2 December 2027 and AI embedded in regulated products to 2 August 2028, while adding a new prohibition on AI generated non consensual intimate imagery and child sexual abuse material. The general purpose AI rules and the transparency obligations for AI generated content were left on schedule.⁹⁴ The architecture, in other words, survived intact; what moved was the timetable. For humanitarian actors the delay matters directly, because the migration and asylum category of Annex III, the part of the Act closest to their work, now takes full effect at the end of 2027 rather than in 2026.

12.3 The United States: A Different Theory of Regulation

The United States has no comprehensive federal AI statute, and the absence is a choice, not an oversight. American AI governance is made of three moving parts: executive orders that change with administrations, sectoral enforcement by existing agencies, and an increasingly dense layer of state law. The Biden administration's 2023 executive order on safe, secure, and trustworthy AI, with its reporting duties for frontier training runs, was revoked in January 2025 by the incoming administration's order Removing Barriers to American Leadership in Artificial Intelligence, and the policy was recast in July 2025 as America's AI Action Plan, organized around accelerating innovation, building AI infrastructure, and exporting the American AI stack; the AI Safety Institute was refashioned as the Center for AI Standards and Innovation.⁹⁵ Congress, for its part, has legislated only at the edges, most notably the TAKE IT DOWN Act on non consensual intimate imagery, and

⁹² European Parliament and Council of the European Union, *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*, Official Journal of the European Union, 12 July 2024, art. 5 and annex III.

⁹³ European Commission, "AI Act: Application Timeline," Shaping Europe's Digital Future, digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai; Regulation (EU) 2024/1689, art. 113.

⁹⁴ Gibson Dunn, "EU AI Act Omnibus Agreement: Postponed High-Risk Deadlines and Other Key Changes," Gibson Dunn client alert, May 2026; European Commission, Digital Omnibus package, provisional agreement of 6 May 2026.

⁹⁵ Executive Order 14179, "Removing Barriers to American Leadership in Artificial Intelligence," 23 Jan. 2025; The White House, *Winning the Race: America's AI Action Plan*, July 2025.

when the July 2025 budget bill briefly contained a ten year moratorium on state AI enforcement, the Senate stripped it out by a vote of ninety nine to one.⁹⁶

Into that federal vacuum the states have moved. More than a thousand AI bills were introduced in state legislatures in 2025 alone, and on 1 January 2026 a cluster of significant laws took effect: California’s Transparency in Frontier Artificial Intelligence Act, which requires the largest frontier developers to publish catastrophic risk frameworks and report critical safety incidents; California’s training data disclosure law; Texas’s Responsible Artificial Intelligence Governance Act, which prohibits development and deployment for defined harmful purposes; and Illinois’s prohibition on AI driven employment discrimination. Colorado’s pioneering comprehensive statute, the first in the nation, saw its enforcement date pushed to 30 June 2026 while the legislature narrows its scope.⁹⁷ The federal response arrived on 11 December 2025 as an executive order, Ensuring a National Policy Framework for Artificial Intelligence, which established an AI Litigation Task Force to challenge state laws in court, directed the Commerce Department to catalogue state statutes deemed inconsistent with federal policy, and conditioned certain federal funds on states refraining from onerous AI regulation. California’s governor answered with an executive order of his own directing state agencies to develop AI safety requirements. The result, as of mid 2026, is an open constitutional contest between federal deregulation and state legislation.⁹⁸

12.4 Two Regulatory Worlds: How They Differ

Dimension	European Union	United States
Legal form	One binding horizontal regulation for the whole market	No comprehensive statute; executive orders, agency action, and state laws
Regulatory theory	Precautionary and rights based; obligations before market entry	Innovation first; enforcement after harm, largely through existing law
Scope	All providers and deployers placing systems on the EU market, wherever based	Targeted slices: frontier developers in California, harmful uses in Texas, hiring in Illinois
Treatment of risk	Risk tiers fixed in law, with outright prohibitions	Defined by each instrument; almost no outright federal prohibitions
Enforcement	AI Office and national authorities; fines up to seven percent of global turnover	Agencies, state attorneys general, and private litigation

⁹⁶ TAKE IT DOWN Act, Public Law 119-12, 19 May 2025; “US Senate Strikes AI Regulation Moratorium from Trump Megabill,” Reuters, 1 July 2025.

⁹⁷ California Transparency in Frontier Artificial Intelligence Act (SB 53, 2025); California AB 2013 (2024); Texas Responsible Artificial Intelligence Governance Act (2025); Illinois HB 3773 (2024); Colorado SB 24-205 (2024), enforcement postponed to 30 June 2026. See King and Spalding, “New State AI Laws Are Effective on January 1, 2026, But a New Executive Order Signals Disruption,” Jan. 2026.

⁹⁸ Executive Order, “Ensuring a National Policy Framework for Artificial Intelligence,” 11 Dec. 2025; Cooley, “State AI Laws: Where Are They Now?,” 24 Apr. 2026; California Executive Order N-5-26, 30 Mar. 2026.

Dimension	European Union	United States
Trajectory in 2026	Deadlines deferred to 2027 and 2028; architecture intact	Federal push to preempt states; states legislating anyway

The differences flow from different answers to the question this paper opened with: who should govern the power that AI creates. The European answer, in Bradford’s terms rights driven, treats AI as a regulated product whose risks must be addressed before deployment, and accepts the cost in speed. The American answer, market driven, treats premature regulation as the greater risk, bets on innovation and existing law, and tolerates a fifty state patchwork as the price of federal restraint.⁹⁹ Yet the two worlds are converging in places: both now prohibit AI generated intimate imagery abuse, both single out frontier models for special attention, and both spent 2026 recalibrating under competitiveness pressure, the EU by deferring its deadlines, the United States by attacking its own states’ rules. For the humanitarian sector the practical consequence is asymmetry: a refugee status algorithm deployed in Europe will face binding high risk obligations from the end of 2027, while the same system in the United States faces no dedicated federal rule at all, and its treatment depends on the state in which it operates. The landscape, in short, is not one of regulation versus none, but of one jurisdiction building a single floor slowly and the other fighting over whether floors should exist.

12.5 The Geopolitics of Competition

Against this regulatory current runs the geopolitics of competition. The United States’ October 2022 semiconductor export controls marked, in Gregory Allen’s phrase, a strategy of choking off China’s access to the future of AI, a shift from preserving relative advantage to actively degrading a rival’s capability.¹⁰⁰ Scholars of compute governance argue that computing power has become the most governable input to AI, detectable, excludable, and concentrated in a narrow supply chain, making chips the decisive lever of AI policy, with all the risks of concentration that implies.¹⁰¹ Paul Scharre frames the resulting contest as a struggle across four battlegrounds, data, compute, talent, and institutions, in which AI reshapes military and geopolitical power itself.¹⁰²

12.6 A Fragile Multilateralism

The multilateral response has been real but fragile. The Bletchley Declaration of November 2023 achieved a joint acknowledgment by the United States, China, the European Union, and twenty five other states that frontier AI carries the potential for catastrophic harm, and launched a summit series that continued in Seoul in 2024. By the Paris AI Action Summit of February 2025, the consensus had visibly fractured: the United States and the United

⁹⁹ Anu Bradford, *Digital Empires: The Global Battle to Regulate Technology*, Oxford UP, 2023.

¹⁰⁰ Gregory C. Allen, “Choking Off China’s Access to the Future of AI,” Center for Strategic and International Studies, 11 Oct. 2022.

¹⁰¹ Girish Sastry et al., “Computing Power and the Governance of Artificial Intelligence,” arXiv:2402.08797, Feb. 2024.

¹⁰² Paul Scharre, *Four Battlegrounds: Power in the Age of Artificial Intelligence*, W. W. Norton, 2023.

Kingdom declined to sign the summit statement, and the framing shifted from safety to action and innovation.¹⁰³ Scholars describe the resulting landscape as a weak regime complex, a fragmented array of overlapping initiatives in which, as Huw Roberts and colleagues put it, “siloefforts have led to detrimental fragmentation.”¹⁰⁴ The UN’s High level Advisory Body quantified the exclusion built into this landscape: at the time of its 2024 report, only seven countries were parties to all of the major AI governance initiatives it surveyed, while 118 countries, overwhelmingly in the Global South, were parties to none.¹⁰⁵

12.7 AI at War

War has sharpened every one of these debates. The ICRC has called on states to adopt new legally binding rules on autonomous weapon systems, including a prohibition on systems designed or used to apply force directly against persons.¹⁰⁶ Reports that an AI decision support system known as Lavender was used to generate targeting recommendations in Gaza have produced the first sustained legal scholarship on AI assisted targeting under international humanitarian law, testing the law of distinction, proportionality, and precaution against machine speed war.¹⁰⁷ For humanitarian organizations these developments are not abstract. The same conflict zones where AI now assists targeting are the zones where aid agencies deploy their own algorithms, and where the neutrality and independence of humanitarian data infrastructure is most easily compromised. The war in Ukraine has added a further dimension: the displacement it produced became a hunting ground for technology facilitated human trafficking, a threat the CoLab section returns to, while drone warfare and AI enabled targeting accelerated the militarization of the very technologies humanitarians hope to repurpose for protection.

For the humanitarian sector, the international affairs context yields a sobering conclusion. The rules governing AI are being written by and for the powerful, in fora where affected populations, and often the states that host them, have no seat. Rachel Adams and others argue that AI is thereby deepening global inequality along familiar colonial lines.¹⁰⁸ Humanitarian AI governance must therefore be understood as, among other things, a project of representation: an attempt to force the interests of the excluded into rooms where they are otherwise absent.

¹⁰³ “The Bletchley Declaration,” GOV.UK, 1 Nov. 2023; “Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet,” AI Action Summit, Paris, 11 Feb. 2025.

¹⁰⁴ Huw Roberts et al., “Global AI Governance: Barriers and Pathways Forward,” *International Affairs*, vol. 100, no. 3, 2024, pp. 1275 to 1286.

¹⁰⁵ United Nations High-level Advisory Body on Artificial Intelligence, *Governing AI for Humanity: Final Report*, United Nations, Sept. 2024.

¹⁰⁶ International Committee of the Red Cross, “ICRC Position on Autonomous Weapon Systems,” ICRC, 12 May 2021.

¹⁰⁷ Emelie Andersin, “The Use of the ‘Lavender’ in Gaza and the Law of Targeting: AI-Decision Support Systems and Facial Recognition Technology,” *Journal of International Humanitarian Legal Studies*, vol. 16, no. 2, 2025, pp. 336 to 370.

¹⁰⁸ Rachel Adams, *The New Empire of AI: The Future of Global Inequality*, Polity, 2025.

13. Human Rights Foundations

13.1 The Binding Floor

If the geopolitics of AI is a story of fragmentation, human rights law offers the opposite: a nearly universal, binding normative floor that predates the technology and does not need to be reinvented for it. The Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights, and the International Covenant on Economic, Social and Cultural Rights already guarantee privacy, non discrimination, social security, food, and health, every one of which is implicated when an algorithm allocates aid, flags an asylum claim, or scores vulnerability.¹⁰⁹

13.2 The UN Machinery

The UN human rights machinery grasped this early and concretely. Philip Alston's 2019 report as Special Rapporteur on extreme poverty remains the canonical warning. Examining the digitization of welfare systems worldwide, he described private vendors operating in a human rights free zone and cautioned against "the grave risk of stumbling, zombie-like, into a digital welfare dystopia."¹¹⁰ The digital welfare state is the closest domestic analogue to humanitarian AI, algorithmic systems deciding who receives the means of survival, and its documented failures are a standing caution for aid agencies. A year later, E. Tendayi Achiume, Special Rapporteur on contemporary forms of racism, delivered the first systematic UN analysis of how emerging digital technologies produce structurally racialized outcomes: "Technology is not neutral or objective," she wrote. "It is fundamentally shaped by the racial, ethnic, gender and other inequalities prevalent in society, and typically makes these inequalities worse."¹¹¹ The Office of the High Commissioner for Human Rights went further still in 2021, calling for a moratorium on the sale and use of AI systems that pose serious risks to human rights until adequate safeguards are in place.¹¹²

13.3 Scholarship and Its Critics

Scholars have built a rigorous framework on this doctrinal foundation. Lorna McGregor, Daragh Murray, and Vivian Ng argue that international human rights law supplies what scattered AI ethics cannot: an end to end accountability framework spanning the full

¹⁰⁹ United Nations General Assembly, *Universal Declaration of Human Rights*; *International Covenant on Civil and Political Rights*, 16 Dec. 1966, art. 17; *International Covenant on Economic, Social and Cultural Rights*, 16 Dec. 1966, arts. 9, 11, and 12.

¹¹⁰ Philip Alston, *Digital Welfare States and Human Rights: Report of the Special Rapporteur on Extreme Poverty and Human Rights*, UN Doc. A/74/493, 11 Oct. 2019.

¹¹¹ E. Tendayi Achiume, *Racial Discrimination and Emerging Digital Technologies: A Human Rights Analysis*, UN Doc. A/HRC/44/57, June 2020; Office of the UN High Commissioner for Human Rights, "Emerging Digital Technologies Entrench Racial Inequality, UN Expert Warns," press release, 15 July 2020.

¹¹² United Nations High Commissioner for Human Rights, *The Right to Privacy in the Digital Age*, UN Doc. A/HRC/48/31, Sept. 2021.

algorithmic lifecycle, with defined harms, state obligations, and business responsibilities.¹¹³ Karen Yeung and colleagues put the point polemically: voluntary ethics frameworks amount to ethics washing, and only mandatory, human rights centred design, deliberation, and oversight backed by law can end it.¹¹⁴ Mathias Risse has mapped the philosophical terrain, urging the human rights community to engage AI across time horizons, discrimination and exclusion now, concentration of power later.¹¹⁵ Alessandro Mantelero has translated the framework into method, developing the human rights, ethical, and social impact assessment model that influenced the Council of Europe's own assessment methodology.¹¹⁶ The UN Guiding Principles on Business and Human Rights supply the corporate half of the equation. The OHCHR's B-Tech project has applied their due diligence architecture directly to AI companies, concluding that "the most significant harms to people related to generative AI are in fact impacts on internationally agreed human rights."¹¹⁷

The human rights approach has honest internal critics. Nathalie Smuha, sympathetic to the project, warns that human rights are a necessary floor but not a sufficient ceiling: without concretizing legislation, enforcement capacity, and a supporting societal infrastructure, any human rights based governance framework for AI risks falling short of its purpose.¹¹⁸ The point is well taken, and it is precisely why the humanitarian sector matters to this debate. Humanitarian organizations cannot wait for enforcement infrastructure to mature; they operate today, in jurisdictions where none exists. For them, the human rights framework functions less as enforceable law than as professional ethic and design constraint, which is why the practitioner tooling for human rights impact assessment of digital activities developed by the Danish Institute for Human Rights and others has been so readily absorbed into humanitarian practice.¹¹⁹

14. The UN System's Evolving Approach to AI

The United Nations came to AI along two tracks, using it and governing it, and the story of the past decade is the gradual convergence of the two.

¹¹³ Lorna McGregor, Daragh Murray, and Vivian Ng, "International Human Rights Law as a Framework for Algorithmic Accountability," *International and Comparative Law Quarterly*, vol. 68, no. 2, 2019, pp. 309 to 343.

¹¹⁴ Karen Yeung, Andrew Howes, and Ganna Pogrebna, "AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing," *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, Oxford UP, 2020, pp. 77 to 106.

¹¹⁵ Mathias Risse, "Human Rights and Artificial Intelligence: An Urgently Needed Agenda," *Human Rights Quarterly*, vol. 41, no. 1, 2019, pp. 1 to 16.

¹¹⁶ Alessandro Mantelero, *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI*, T.M.C. Asser Press, 2022.

¹¹⁷ Office of the United Nations High Commissioner for Human Rights, *Guiding Principles on Business and Human Rights*, United Nations, 2011; OHCHR B-Tech Project, *Taxonomy of Human Rights Risks Connected to Generative AI*, OHCHR, Nov. 2023.

¹¹⁸ Nathalie A. Smuha, "Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea," *Philosophy and Technology*, vol. 34, no. 1, 2021, pp. 91 to 104.

¹¹⁹ The Danish Institute for Human Rights, *Guidance on Human Rights Impact Assessment of Digital Activities*, DIHR, 2020.

14.1 The Usage Track

The usage track began early. Global Pulse made the UN an early adopter of big data analytics for development and crisis monitoring. The International Telecommunication Union's AI for Good Global Summit, launched in 2017 with dozens of UN agencies as partners, became the system's flagship AI platform, and the ITU's annual surveys of United Nations activities on artificial intelligence now catalogue hundreds of AI projects across the system.¹²⁰ Individual agencies developed their own instruments. UNICEF's policy guidance on AI for children set out nine requirements for child centred AI.¹²¹ UNHCR embedded a rights based AI approach in its digital transformation strategy, and WFP co led the drafting of system wide AI principles while operating HungerMap LIVE.¹²² OCHA's Centre for Humanitarian Data built the peer review framework for predictive analytics and, in the generative era, issued briefing notes on AI in the humanitarian sector.¹²³ UNDP launched an AI Hub for Sustainable Development to strengthen local AI ecosystems in developing economies.¹²⁴ Internally, the Secretary General's UN 2.0 agenda made data, digital, and innovation capabilities a matter of institutional reform.¹²⁵

14.2 The Governance Track

The governance track unfolded through a deliberate sequence. The High level Panel on Digital Cooperation delivered *The Age of Digital Interdependence* in 2019, seeding the architecture that followed.¹²⁶ The Secretary General's Roadmap for Digital Cooperation committed the UN in 2020 to global cooperation on AI that is trustworthy, human rights based, safe, and sustainable.¹²⁷ UNESCO's Recommendation, examined above, became the first global normative instrument on AI ethics in 2021. In 2022 the UN turned the lens inward: the High level Committee on Programmes adopted ten Principles for the Ethical Use of Artificial Intelligence in the United Nations System, from do no harm to inclusion and participation, binding the system's own operations to the standards it urges on others.¹²⁸

¹²⁰ International Telecommunication Union, "AI for Good Global Summit," ITU, aiforgood.itu.int; International Telecommunication Union, *United Nations Activities on Artificial Intelligence*, ITU, annual editions.

¹²¹ UNICEF, *Policy Guidance on AI for Children 2.0*, UNICEF Office of Global Insight and Policy, Nov. 2021.

¹²² UNHCR, "Our Approach to Artificial Intelligence," *UNHCR Digital Transformation Strategy 2022 to 2026*, UNHCR; World Food Programme Innovation Accelerator, "New Global Principles for Innovative and Ethical AI," WFP, 2022.

¹²³ Centre for Humanitarian Data, *Peer Review Framework*; UN OCHA, *Briefing Note on Artificial Intelligence and the Humanitarian Sector*, OCHA, 2024.

¹²⁴ United Nations Development Programme, *AI Hub for Sustainable Development*, UNDP, 2024.

¹²⁵ United Nations Secretary-General, *Our Common Agenda Policy Brief 11: UN 2.0*, United Nations, Sept. 2023.

¹²⁶ UN Secretary-General's High-level Panel on Digital Cooperation, *The Age of Digital Interdependence*, United Nations, June 2019.

¹²⁷ United Nations Secretary-General, *Roadmap for Digital Cooperation*, UN Doc. A/74/821, June 2020.

¹²⁸ High-level Committee on Programmes, *Principles for the Ethical Use of Artificial Intelligence in the United Nations System*, UN System Chief Executives Board for Coordination, Sept. 2022.

14.3 The Political Breakthrough

Then came the political breakthrough. In March 2024 the General Assembly adopted by consensus its first resolution on AI, led by the United States with more than 120 co sponsors, linking safe, secure, and trustworthy AI to sustainable development. In July 2024 a resolution led by China, likewise adopted by consensus and co sponsored by the United States, addressed AI capacity building for developing countries.¹²⁹ The twin resolutions displayed a distinctive dynamic: strategic rivals competing for normative leadership through consensus rather than against it. The Secretary General's High level Advisory Body on AI supplied the intellectual synthesis in its September 2024 report *Governing AI for Humanity*, diagnosing the governance deficit in blunt terms: "No global framework exists to govern AI. With its development in the hands of a few multinational companies in a few countries, the impacts of unleashing AI risk being imposed on most people without their having any say in the decisions for doing so."¹³⁰ The Global Digital Compact, adopted at the Summit of the Future days later, translated diagnosis into commitment, including the creation of an Independent International Scientific Panel on AI and a Global Dialogue on AI Governance.¹³¹

14.4 The 2026 Institutions

Resolution 79/325 of August 2025 gave both mechanisms their terms of reference, and 2026 brought them to life. A forty member panel drawn from all five UN regions, selected from more than 2,600 applications, held its inaugural meeting in March 2026 and elected Yoshua Bengio and Maria Ressa as co chairs. Its first annual report appeared on 1 July 2026, and the first Global Dialogue on AI Governance convened in Geneva the following week.¹³² The symbolism of the co chairs is worth pausing on: one of the world's most cited AI scientists, who also chairs the International AI Safety Report, and a Nobel laureate journalist whose warning to the Dialogue, that without the ability to tell fact from fiction there can be no democracy, anchored AI governance in information integrity.¹³³ The safety discourse born at Bletchley has migrated into the UN system, where it now shares a room with the development, human rights, and humanitarian agendas. Secretary General Guterres has framed the stakes in terms humanitarian actors will recognize: "A world of AI haves and have-nots would be a world of perpetual instability."¹³⁴

¹²⁹ United Nations General Assembly, Resolution A/RES/78/265, *Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development*, 21 Mar. 2024; United Nations General Assembly, Resolution A/RES/78/311, *Enhancing International Cooperation on Capacity-Building of Artificial Intelligence*, 1 July 2024.

¹³⁰ United Nations High-level Advisory Body on Artificial Intelligence, *Governing AI for Humanity*, press release, 19 Sept. 2024.

¹³¹ United Nations, *Global Digital Compact*, annex to Resolution A/RES/79/1, 22 Sept. 2024, objective 5.

¹³² United Nations General Assembly, Resolution A/RES/79/325, 26 Aug. 2025; United Nations, "Independent International Scientific Panel on AI: FAQ," United Nations, 2026.

¹³³ "Global Push for AI Governance amid Warnings of 'Catastrophic Harm,'" *UN News*, 5 July 2026.

¹³⁴ António Guterres, quoted in United Nations, "Global Digital Compact," un.org/global-digital-compact/en.

The UN's approach remains more scaffolding than edifice. The Scientific Panel assesses but does not regulate; the Global Dialogue convenes but does not bind. Yet for the humanitarian sector the trajectory matters. For the first time, the institutions that coordinate humanitarian response, the norms that protect human rights, and the fora that govern AI sit within a single system: imperfect, under resourced, and indispensable.

15. The Ethical Tech CoLab: Motivation and Research Philosophy

15.1 Motivation

The Ethical Tech CoLab is a research initiative of New York University's School of Professional Studies Center for Global Affairs, conducted in collaboration with Microsoft from 2024 to 2026. Its mission is to explore intervention opportunities at the intersection of emerging technologies and the human condition.¹³⁵ The CoLab sits at the meeting point of the two worlds this paper has described: the academic study of ethics in international affairs and the practical governance of emerging technology. Its founding observation is the one with which this paper began. The problem is power; ethics is the answer. Technology companies possess capabilities that reshape the lives of people who never chose them, and the populations most affected, refugees, trafficking survivors, communities in conflict, are precisely those least represented in the rooms where technological choices are made.

The CoLab's motivation is therefore corrective. It exists to bring the deliberative tradition described in the opening sections of this paper into direct contact with the technologies and governance regimes described in the later ones, and to do so from within the institutions that build and govern them rather than at a critical distance alone. The collaboration between a university and a technology company is itself a methodological statement. The academy contributes independence, historical depth, and the pluralism of ethical frameworks; industry contributes operational knowledge of how systems are actually built, deployed, and misused. Neither alone is sufficient. The partnership is exemplified by practitioners who hold both roles at once, such as Microsoft's director of digital transformation, who also teaches emerging technology and ethical sourcing at the Center for Global Affairs, and whose work on technology facilitated human trafficking demonstrates the model in practice.¹³⁶

15.2 The Portfolio Through Three Lenses

The CoLab's published portfolio expresses this orientation concretely, and each project is best understood through three lenses: the humanitarian problem it addresses, the ethical

¹³⁵ NYU Ethical Tech CoLab, "Publications," NYU School of Professional Studies Center for Global Affairs and Microsoft, ethical-tech-colab.github.io/website/publications.

¹³⁶ Author's field notes, "Developing Guidelines on the Intersection of Human Trafficking and Technology," working group meeting hosted by the Libertas Council at the European Parliament, Brussels, 3 Dec. 2024. Speaker affiliations as listed in the meeting agenda.

safeguards it incorporates, and the measurable public value it seeks to create.¹³⁷ The largest body of work concerns civilian evacuation in armed conflict. The Civilian Evacuation Risk Anticipation Index is a decision support tool for anticipating risks to civilians who must decide whether and when to flee; the Evacuation Simulator models civilian departure decisions through agent based simulation; the Mariupol Corridor Severity Model tracks the changing danger of the evacuation corridors from the besieged city of Mariupol between March and May of 2022; and After the Corridor connects evacuation response to the longer arc of refugee economic inclusion. The humanitarian problem is the most consequential decision a civilian in a war zone ever makes, taken under conditions of terror and informational collapse. The safeguard is that these are decision support instruments for human judgment, not automated directives, and that they are built on documented, contestable data. The public value is measurable in the currency that matters most: earlier, better informed protection decisions.

A second cluster addresses exploitation and traceability. The Forced Labor Structural Risk Index assesses, at the country level, the structural conditions under which forced labor flourishes. It belongs to the same fight documented in this author's fieldwork at the European Parliament, where working groups convened in December 2024 to develop guidelines on the intersection of human trafficking and technology heard how traffickers have sharpened their use of technological advances to recruit, groom, coerce, advertise, and sell victims online, with refugee and migrant women displaced by the war in Ukraine targeted through the very messaging platforms on which they depend for survival information.¹³⁸ The OSCE's Special Representative for Combating Trafficking in Human Beings described mapping thousands of online commercial sex websites across participating states and warned that artificial intelligence is expanding traffickers' reach through automated translation, targeted false advertisements, and synthetic media used for coercion, while the same technologies, responsibly governed, could help investigators detect patterns and reach victims earlier.¹³⁹ The International Labour Organization estimates the annual illegal profits from forced labour at 236 billion United States dollars, a figure that locates this harm among the largest criminal economies the technology sector inadvertently hosts.¹⁴⁰ Further projects extend the same logic to other domains of the human condition: the Digital Provenance Passport and related provenance tools bring automated transparency to the ownership histories of cultural property; HASTE applies

¹³⁷ NYU Ethical Tech CoLab, "Publications." Project descriptions in this and the following two paragraphs are drawn from the CoLab's published reports, including *The Civilian Evacuation Risk Anticipation Index*, *The Evacuation Simulator*, *The Mariupol Corridor Severity Model*, *After the Corridor* (July 2026), *The Forced Labor Structural Risk Index* (Dec. 2025), *The Digital Provenance Passport*, *HASTE: High-speed Assessment and Satellite Tracking for Emergencies*, *The Diplomatic Simulator* (July 2026), and *AI's Carbon Footprint* (May 2025).

¹³⁸ Libertas Council, "Developing Guidelines on the Intersection of Human Trafficking and Technology," meeting concept note and proceedings, European Parliament, 3 Dec. 2024; author's field notes.

¹³⁹ Remarks of Kari Johnstone, OSCE Special Representative and Co-ordinator for Combating Trafficking in Human Beings, recorded in author's field notes, 3 Dec. 2024; see also OSCE Office of the Special Representative and Co-ordinator for Combating Trafficking in Human Beings and Tech Against Trafficking, *Leveraging Innovation to Fight Trafficking in Human Beings: A Comprehensive Analysis of Technology Tools*, OSCE, 2020.

¹⁴⁰ International Labour Organization, *Profits and Poverty: The Economics of Forced Labour*, 2nd ed., ILO, Mar. 2024.

satellite based assessment to disaster response; the Diplomatic Simulator models multi party negotiation; and the CoLab’s research on AI’s carbon footprint turns the ethical lens on the technology’s own environmental costs.¹⁴¹

15.3 Research Philosophy

Figure 4. The five pillars of ethics of the Ethical Tech CoLab

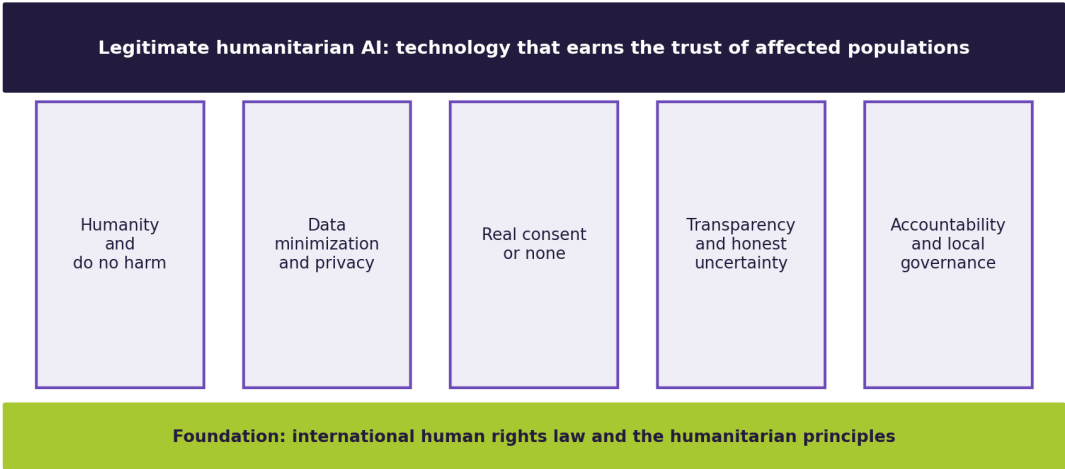


Figure 2. The five pillars of ethics of the Ethical Tech CoLab.

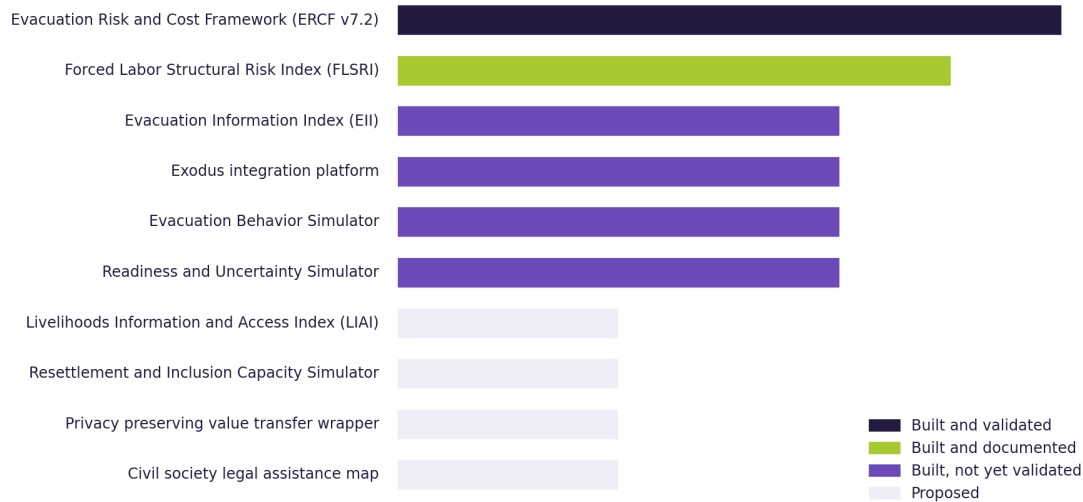
Three commitments define the research philosophy. The first is ethics before algorithms. As this paper has argued, the CoLab treats the current wave of artificial intelligence not as an unprecedented rupture but as the latest chapter in a four thousand year effort to govern power, and it insists on understanding what ethics is before prescribing what AI should do. This is why the CoLab’s research begins from moral philosophy, international relations, and humanitarian ethics rather than from model architectures. The second is ethical deliberation as method. Following the Carnegie Council’s framework, the CoLab approaches each problem by identifying the ethical stakes, deliberating across plural frameworks, acting, and taking responsibility for outcomes, with the deliberation conducted wherever possible with those most affected rather than about them.¹⁴² The third is the human rights floor. Consistent with the UNESCO Recommendation, the UN system’s own principles for ethical AI use, and the human rights scholarship examined above, the CoLab evaluates every proposed technology against internationally agreed rights rather than against corporate or institutional convenience, and it treats the humanitarian do no harm obligation as a hard constraint rather than an optimization target.¹⁴³

¹⁴¹ NYU Ethical Tech CoLab, “Publications.”

¹⁴² Carnegie Council, *Ethics in International Affairs* 101, pp. 12 to 15.

¹⁴³ UNESCO, *Recommendation on the Ethics of Artificial Intelligence*; High-level Committee on Programmes, *Principles for the Ethical Use of Artificial Intelligence in the United Nations System*.

Figure 1. Maturity of the Ethical Tech CoLab portfolio, July 2026



Maturity registers as stated in Sections 3 and 13 of the report.

Figure 3. Maturity of the Ethical Tech CoLab portfolio, July 2026.

The CoLab’s ambition, of which this paper is a first expression, is a body of research useful to policymakers, humanitarian practitioners, donors, and researchers who must make decisions about artificial intelligence without becoming AI engineers. Its wager is that the disciplines this paper has drawn together, philosophy, ethics, international affairs, humanitarian ethics, digital ethics, and responsible innovation, are not six literatures but one conversation, and that the institutions that learn to hold that conversation well will be the ones that earn the trust on which the legitimate use of artificial intelligence depends.

16. Conclusion

This paper set out to answer a question that is asked constantly and defined rarely: what is ethical AI? The answer it has defended is institutional rather than technical. Ethical AI is artificial intelligence governed by the same disciplines humanity has always required of power: deliberation across plural values, accountability to those affected, a floor of universal rights, and the humility to take responsibility when judgment fails. Ethics did not emerge because human beings became virtuous. It emerged because societies learned that power without restraint destroys the communities on which power depends. Artificial intelligence does not change that lesson. It makes the lesson urgent.

The humanitarian sector is where the urgency is greatest and the test is hardest. There, consent collapses, experimentation finds the least protected subjects, accountability runs to people with no vote and no purchase, and the cost of failure is measured in lives. A technology that can be governed ethically there can be governed ethically anywhere. The frameworks now taking shape, from UNESCO’s Recommendation to the UN’s Scientific Panel and Global Dialogue, from the EU AI Act to the ICRC’s precautionary policy, are the scaffolding of that governance. Whether they become an edifice depends on whether

institutions, including universities, technology companies, and the collaborations between them, do the slow deliberative work this paper has tried to model. Societies do not flourish because they possess increasingly powerful tools. They flourish because they develop increasingly trustworthy ways of governing them. That has been the purpose of ethics for four thousand years. It is the purpose of ethical AI now.

Works Cited

- Achiume, E. Tendayi. *Racial Discrimination and Emerging Digital Technologies: A Human Rights Analysis. Report of the Special Rapporteur on Contemporary Forms of Racism.* UN Doc. A/HRC/44/57, United Nations, June 2020.
- Acton, John Emerich Edward Dalberg. Letter to Mandell Creighton, 5 Apr. 1887. *Historical Essays and Studies*, edited by John Neville Figgis and Reginald Vere Laurence, Macmillan, 1907.
- Adams, Rachel. *The New Empire of AI: The Future of Global Inequality.* Polity, 2025.
- Allen, Gregory C. “Choking Off China’s Access to the Future of AI.” Center for Strategic and International Studies, 11 Oct. 2022, [csis.org/analysis/choking-chinas-access-future-ai](https://www.csis.org/analysis/choking-chinas-access-future-ai).
- Alston, Philip. *Digital Welfare States and Human Rights: Report of the Special Rapporteur on Extreme Poverty and Human Rights.* UN Doc. A/74/493, United Nations, 11 Oct. 2019.
- Amnesty International and Access Now. *The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems.* 16 May 2018.
- Andersin, Emelie. “The Use of the ‘Lavender’ in Gaza and the Law of Targeting: AI-Decision Support Systems and Facial Recognition Technology.” *Journal of International Humanitarian Legal Studies*, vol. 16, no. 2, 2025, pp. 336 to 370.
- Appiah, Kwame Anthony. *Cosmopolitanism: Ethics in a World of Strangers.* W. W. Norton, 2006.
- Aristotle. *Nicomachean Ethics.* Translated by Terence Irwin, 2nd ed., Hackett, 1999.
- Beduschi, Ana. “Harnessing the Potential of Artificial Intelligence for Humanitarian Action: Opportunities and Risks.” *International Review of the Red Cross*, vol. 104, no. 919, 2022, pp. 1149 to 1169.
- Bengio, Yoshua, et al. *International AI Safety Report 2026.* AI Security Institute, Feb. 2026, internationalaisafetyreport.org.
- Benjamin, Ruha. *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity, 2019.

- Bietti, Elettra. "From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM, 2020, pp. 210 to 219.
- "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1 and 2 November 2023." GOV.UK, 1 Nov. 2023, gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration.
- Bradford, Anu. *Digital Empires: The Global Battle to Regulate Technology*. Oxford UP, 2023.
- Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research*, vol. 81, 2018, pp. 77 to 91.
- Carnegie Council for Ethics in International Affairs. *Ethics in International Affairs 101*. Carnegie Council, carnegiecouncil.org.
- Centre for Humanitarian Data. *Peer Review Framework for Predictive Analytics in Humanitarian Response*. UN OCHA, Mar. 2020.
- The Code of Hammurabi*. Translated by L. W. King, 1910. The Avalon Project, Yale Law School, avalon.law.yale.edu/ancient/hamframe.asp.
- Confucius. *Analects*. Translated by Edward Slingerland, Hackett, 2003.
- Coppi, Giulio, Rebeca Moreno Jimenez, and Sofia Kyriazi. "Explicability of Humanitarian AI: A Matter of Principles." *Journal of International Humanitarian Action*, vol. 6, art. 19, 2021.
- Cooley. "State AI Laws: Where Are They Now?" Cooley, 24 Apr. 2026.
- Council of Europe. *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*. Council of Europe Treaty Series no. 225, 5 Sept. 2024.
- Crawford, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale UP, 2021.
- The Danish Institute for Human Rights. *Guidance on Human Rights Impact Assessment of Digital Activities*. DIHR, 2020.
- Diakopoulos, Nicholas, et al. "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms." FAT/ML, fatml.org/resources/principles-for-accountable-algorithms.
- Devidal, Pierrick. "Lost in Digital Translation? The Humanitarian Principles in the Digital Age." *International Review of the Red Cross*, vol. 106, no. 925, 2024.
- The Engine Room and Oxfam. *Biometrics in the Humanitarian Sector*. The Engine Room, 2018, updated 2023.

- European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. Official Journal of the European Union, 12 July 2024.
- Friedman, Batya, and David G. Hendry. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, 2019.
- Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*. Berkman Klein Center Research Publication no. 2020-1, Harvard University, 2020.
- Floridi, Luciano, and Josh Cowls. “A Unified Framework of Five Principles for AI in Society.” *Harvard Data Science Review*, vol. 1, no. 1, 2019.
- Future of Life Institute. “Asilomar AI Principles.” Future of Life Institute, 2017, futureoflife.org/open-letter/ai-principles.
- Gibson Dunn. “EU AI Act Omnibus Agreement: Postponed High-Risk Deadlines and Other Key Changes.” Gibson Dunn, May 2026.
- Geburu, Timnit, et al. “Datasheets for Datasets.” *Communications of the ACM*, vol. 64, no. 12, 2021, pp. 86 to 92.
- “Global Push for AI Governance amid Warnings of ‘Catastrophic Harm.’” *UN News*, 5 July 2026, news.un.org/en/story/2026/07/1167862.
- Heinzelman, Jessica, and Carol Waters. *Crowdsourcing Crisis Information in Disaster-Affected Haiti*. Special Report 252, United States Institute of Peace, 2010.
- High-level Committee on Programmes, UN System Chief Executives Board for Coordination. *Principles for the Ethical Use of Artificial Intelligence in the United Nations System*. United Nations, Sept. 2022.
- High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. European Commission, 8 Apr. 2019.
- Hobbes, Thomas. *Leviathan*. Edited by Richard Tuck, Cambridge UP, 1996.
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. 1st ed., IEEE, 2019.
- Inter-Agency Standing Committee. *IASC Operational Guidance on Data Responsibility in Humanitarian Action*. IASC, 2023.
- International Committee of the Red Cross. *Building a Responsible Humanitarian Approach: The ICRC’s Policy on Artificial Intelligence*. ICRC, Nov. 2024.
- International Committee of the Red Cross. “ICRC Position on Autonomous Weapon Systems.” ICRC, 12 May 2021.

- International Labour Organization. *Profits and Poverty: The Economics of Forced Labour*. 2nd ed., ILO, Mar. 2024.
- International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC 42001:2023, Information Technology, Artificial Intelligence, Management System*. ISO, Dec. 2023.
- International Telecommunication Union. *United Nations Activities on Artificial Intelligence*. ITU, annual editions.
- Jacobsen, Katja Lindskov. "Experimentation in Humanitarian Locations: UNHCR and Biometric Registration of Afghan Refugees." *Security Dialogue*, vol. 46, no. 2, 2015, pp. 144 to 164.
- Jobin, Anna, Marcello Ienca, and Effy Vayena. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence*, vol. 1, 2019, pp. 389 to 399.
- Johnson, Emily. "AI Ethics Frameworks: 10 Essential Resources to Build an Ethical AI Framework." *SecureITWorld*, 12 Mar. 2025, secureitworld.com/blog/ai-ethics-frameworks-10-essential-resources-to-build-an-ethical-ai-framework.
- Juskalian, Russ. "Inside the Jordan Refugee Camp That Runs on Blockchain." *MIT Technology Review*, 12 Apr. 2018.
- Kant, Immanuel. "Toward Perpetual Peace." *Practical Philosophy*, translated and edited by Mary J. Gregor, Cambridge UP, 1996.
- Karenga, Maulana. *Maat, the Moral Ideal in Ancient Egypt: A Study in Classical African Ethics*. Routledge, 2004.
- King and Spalding. "New State AI Laws Are Effective on January 1, 2026, But a New Executive Order Signals Disruption." King and Spalding, Jan. 2026.
- Kuner, Christopher, and Massimo Marelli, editors. *Handbook on Data Protection in Humanitarian Action*. 2nd ed., ICRC and Brussels Privacy Hub, 2020.
- Laidlaw, James. *The Subject of Virtue: An Anthropology of Ethics and Freedom*. Cambridge UP, 2014.
- Latonero, Mark. "Stop Surveillance Humanitarianism." *The New York Times*, 11 July 2019.
- Libertas Council. "Developing Guidelines on the Intersection of Human Trafficking and Technology." Working group meeting, European Parliament, Brussels, 3 Dec. 2024.
- Madianou, Mirca. "Technocolonialism: Digital Innovation and Data Practices in the Humanitarian Response to Refugee Crises." *Social Media + Society*, vol. 5, no. 3, 2019, pp. 1 to 13.
- Madianou, Mirca. *Technocolonialism: When Technology for Good Is Harmful*. Polity, 2024.

- Mantelero, Alessandro. *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI*. T.M.C. Asser Press, 2022.
- McGregor, Lorna, Daragh Murray, and Vivian Ng. "International Human Rights Law as a Framework for Algorithmic Accountability." *International and Comparative Law Quarterly*, vol. 68, no. 2, 2019, pp. 309 to 343.
- Meier, Patrick. *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response*. CRC Press, 2015.
- Moor, James H. "What Is Computer Ethics?" *Metaphilosophy*, vol. 16, no. 4, 1985, pp. 266 to 275.
- Mitchell, Margaret, et al. "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2019, pp. 220 to 229.
- Mittelstadt, Brent. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence*, vol. 1, 2019, pp. 501 to 507.
- Morgenthau, Hans J. *Politics Among Nations: The Struggle for Power and Peace*. Alfred A. Knopf, 1948.
- Munro, Robert. "Crowdsourcing and the Crisis-Affected Community: Lessons Learned and Looking Forward from Mission 4636." *Information Retrieval*, vol. 16, no. 2, 2013, pp. 210 to 266.
- National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1, U.S. Department of Commerce, Jan. 2023.
- Nearing, Grey, et al. "Global Prediction of Extreme Floods in Ungauged Watersheds." *Nature*, vol. 627, 2024, pp. 559 to 563.
- Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York UP, 2018.
- North, Douglass C. *Institutions, Institutional Change and Economic Performance*. Cambridge UP, 1990.
- Nussbaum, Martha C. "Patriotism and Cosmopolitanism." *Boston Review*, vol. 19, no. 5, 1994.
- NYU Ethical Tech CoLab. "Publications." NYU School of Professional Studies Center for Global Affairs and Microsoft, ethical-tech-colab.github.io/website/publications.
- Office of the United Nations High Commissioner for Human Rights. *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*. United Nations, 2011.
- Office of the United Nations High Commissioner for Human Rights, B-Tech Project. *Taxonomy of Human Rights Risks Connected to Generative AI*. OHCHR, Nov. 2023.

- Organisation for Economic Co-operation and Development. *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449, OECD, 2019, amended 2024.
- OSCE Office of the Special Representative and Co-ordinator for Combating Trafficking in Human Beings, and Tech Against Trafficking. *Leveraging Innovation to Fight Trafficking in Human Beings: A Comprehensive Analysis of Technology Tools*. OSCE, 2020.
- Ostrom, Elinor. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge UP, 1990.
- Partnership on AI. “Our Tenets.” Partnership on AI, 2016, partnershiponai.org.
- Palen, Leysia, and Kenneth M. Anderson. “Crisis Informatics: New Data for Extraordinary Times.” *Science*, vol. 353, no. 6296, 2016, pp. 224 to 225.
- Pizzi, Michael, Mila Romanoff, and Tim Engelhardt. “AI for Humanitarian Action: Human Rights and Ethics.” *International Review of the Red Cross*, vol. 102, no. 913, 2020, pp. 145 to 180.
- Raji, Inioluwa Deborah, et al. “Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing.” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM, 2020, pp. 33 to 44.
- Russell, Bertrand. *Power: A New Social Analysis*. George Allen and Unwin, 1938.
- Risse, Mathias. “Human Rights and Artificial Intelligence: An Urgently Needed Agenda.” *Human Rights Quarterly*, vol. 41, no. 1, 2019, pp. 1 to 16.
- Roberts, Huw, et al. “Global AI Governance: Barriers and Pathways Forward.” *International Affairs*, vol. 100, no. 3, 2024, pp. 1275 to 1286.
- Sandvik, Kristin Bergtora, Katja Lindskov Jacobsen, and Sean Martin McDonald. “Do No Harm: A Taxonomy of the Challenges of Humanitarian Experimentation.” *International Review of the Red Cross*, vol. 99, no. 904, 2017.
- Sastry, Girish, et al. “Computing Power and the Governance of Artificial Intelligence.” arXiv:2402.08797, Feb. 2024.
- Scharre, Paul. *Four Battlegrounds: Power in the Age of Artificial Intelligence*. W. W. Norton, 2023.
- Signpost AI. “Pilot Report: Signpost AI Information Assistant.” International Rescue Committee, 27 June 2025, signpostai.org/research/pilot-report-signpost-ai-information-assistant.
- Singer, Peter. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton UP, 2011.
- Smuha, Nathalie A. “Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea.” *Philosophy and Technology*, vol. 34, no. 1, 2021, pp. 91 to 104.

- Spencer, Sarah W. *Humanitarian AI: The Hype, the Hope and the Future*. Network Paper 85, Humanitarian Practice Network, ODI, Nov. 2021.
- Spencer, Sarah W. *Humanitarian AI Revisited: Seizing the Potential and Sidestepping the Pitfalls*. Network Paper 89, Humanitarian Practice Network, ODI, May 2024.
- “Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet.” AI Action Summit, Paris, 11 Feb. 2025.
- Stilgoe, Jack, Richard Owen, and Phillip Macnaghten. “Developing a Framework for Responsible Innovation.” *Research Policy*, vol. 42, no. 9, 2013, pp. 1568 to 1580.
- Tomasello, Michael. *A Natural History of Human Morality*. Harvard UP, 2016.
- UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. UNESCO, 23 Nov. 2021.
- UNHCR. “Our Approach to Artificial Intelligence.” *UNHCR Digital Transformation Strategy 2022 to 2026*, UNHCR, unhcr.org/digitalstrategy/ai-approach.
- UNICEF. *Policy Guidance on AI for Children 2.0*. UNICEF Office of Global Insight and Policy, Nov. 2021.
- United Nations. *Global Digital Compact*. Annex to Resolution A/RES/79/1, United Nations, 22 Sept. 2024.
- United Nations Development Programme. *AI Hub for Sustainable Development: Strengthening Local AI Ecosystems through Collective Action*. UNDP, 2024.
- United Nations General Assembly. *Enhancing International Cooperation on Capacity-Building of Artificial Intelligence*. Resolution A/RES/78/311, 1 July 2024.
- United Nations General Assembly. *International Covenant on Civil and Political Rights*. United Nations, 16 Dec. 1966.
- United Nations General Assembly. *International Covenant on Economic, Social and Cultural Rights*. United Nations, 16 Dec. 1966.
- United Nations General Assembly. Resolution A/RES/79/325, *Terms of Reference and Modalities for the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance*. 26 Aug. 2025.
- United Nations General Assembly. *Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development*. Resolution A/RES/78/265, 21 Mar. 2024.
- United Nations General Assembly. *Universal Declaration of Human Rights*. Resolution 217 A (III), 10 Dec. 1948.
- United Nations High Commissioner for Human Rights. *The Right to Privacy in the Digital Age*. UN Doc. A/HRC/48/31, United Nations, Sept. 2021.

- United Nations High-level Advisory Body on Artificial Intelligence. *Governing AI for Humanity: Final Report*. United Nations, Sept. 2024.
- United Nations Secretary-General. *Our Common Agenda Policy Brief 11: UN 2.0*. United Nations, Sept. 2023.
- United Nations Secretary-General. *Roadmap for Digital Cooperation: Report of the Secretary-General*. UN Doc. A/74/821, United Nations, June 2020.
- UN Global Pulse. *Big Data for Development: Challenges and Opportunities*. United Nations, May 2012.
- UN OCHA. *Briefing Note on Artificial Intelligence and the Humanitarian Sector*. OCHA, 2024.
- UN OCHA. *OCHA Data Responsibility Guidelines*. Centre for Humanitarian Data, Oct. 2021.
- UN Secretary-General's High-level Panel on Digital Cooperation. *The Age of Digital Interdependence*. United Nations, June 2019.
- Université de Montréal. *Montreal Declaration for a Responsible Development of Artificial Intelligence*. 2018, montrealdeclaration-responsibleai.com.
- United States, Executive Office of the President. Executive Order 14179, "Removing Barriers to American Leadership in Artificial Intelligence." 23 Jan. 2025.
- United States, Executive Office of the President. "Ensuring a National Policy Framework for Artificial Intelligence." Executive order, 11 Dec. 2025.
- The White House. *Winning the Race: America's AI Action Plan*. July 2025.
- Veale, Michael, and Frederik Zuiderveen Borgesius. "Demystifying the Draft EU Artificial Intelligence Act." *Computer Law Review International*, vol. 22, no. 4, 2021, pp. 97 to 112.
- Weber, Max. *Economy and Society: An Outline of Interpretive Sociology*. Edited by Guenther Roth and Claus Wittich, U of California P, 1978.
- World Bank. "Famine Action Mechanism (FAM)." World Bank, 2018, worldbank.org/en/programs/famine-early-action-mechanism.
- World Food Programme. "WFP Releases HungerMap LIVE." WFP, 2019.
- World Food Programme Innovation Accelerator. "New Global Principles for Innovative and Ethical AI." WFP, 2022.
- Whittaker, Meredith, et al. *AI Now Report 2018*. AI Now Institute, New York University, Dec. 2018.
- Winner, Langdon. "Do Artifacts Have Politics?" *Daedalus*, vol. 109, no. 1, 1980, pp. 121 to 136.

Yeung, Karen, Andrew Howes, and Ganna Pogrebna. "AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing." *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, Oxford UP, 2020, pp. 77 to 106.