Ethics in Artificial Intelligence
Toward Foundations for Global Policy

Nazli Choucri

Professor
Political Science Department
Massachusetts Institute of Technology

November 24, 2022

Abstract

Matters of ethics are becoming more salient at all levels of politics, almost everywhere. In the scientific community, ethics in AI is increasingly gaining attention. The fact is that the rate of change in AI innovations and applications are growing much faster than our general appreciation or understanding of content or of consequences. There is a large variety of statements, but few ethical practices by countries, corporations, and individuals that are desirable in the ethics domain for the broad area of Artificial intelligence. Occurring far less frequently—if at all—are the operational applications of ethics codes in the innovation, practice, and policy of AI. To date, the focus of attention is on scientific and technical advances, as well as enhanced computational advances.

.

Citation: Choucri, N. (2022). *Ethics in Artificial Intelligence: Toward Foundations for Global Policy* [White Paper]. MIT Political Science Department.

Unique Resource Identifier: https://hdl.handle.net/1721.1/146915

Acknowledgement: This material is based on work supported by the U.S. Department of Defense for National Security Agency, *Science of Security & Privacy Program* under Grant No. H98230-18-D-00-0010. Any opinions, findings, conclusions or recommendations therein are those of the author(s) and do not necessarily reflect the views of the US Department of Defense.

Ethics in Artificial Intelligence
Toward Foundations for Global Policy

1 The Ethics-in-AI Gap

All evidence indicates that the rates of change in innovation and applications of Artificial Intelligence are growing much faster than our ability to fully appreciate the implications or anticipate the consequences thereof. To date, the focus of attention is on scientific and technical advances as well as enhanced computational advances. By contrast, there has been limited systematic attention to attendant human and social issues or to overall societal effects. There are no authoritative reviews of the diverse discourses, disagreements, or disparate statements on social dimensions or normative issues.

Especially important in this connection has been the pervasive "ethics-in-AI gap," that is, the near absence of attention paid to matters of ethics in Artificial Intelligence. At this point, there is a growing recognition that ethical issues cannot be ignored. This view is gaining traction among leading information and communication technology companies central to the AI domain, broadly defined. Some have issued formal statements expressing their corporate position. Others, like Amazon, Google, Facebook, IBM, and Microsoft are collaborating to develop best practices in AI. Governments are gradually turning to these issues as well. Of relevance here are the OECD's AI Policy Observatory and the European Commission's High-level Group on Artificial Intelligence, as well as the United States Commission on Artificial Intelligence.

1.1 Missing Pieces

Missing is a sustained and widely shared scholarly interest in—and systematic attention to—matters of ethics in artificial intelligence, including content, role, implications, or consequences. While several research institutes in the U.S. and Europe are turning to ethical and normative challenges, there appears to also be full-scale research initiatives on, or scientific discussions of, ethical issues amid world-changing technological innovations under conditions of uncertainty, with yet-to-be framed risks to the wellbeing of individuals, societies, and the international system as a whole.

Especially challenging is the absence of a "best" framework for addressing "ethics in AI." For example, consider the difference between a (i) "system of ethics in AI" versus (ii) an "AI system of ethics." Each point to different system referents and boundaries. One refers to ethics *within* the domain of artificial intelligence. The other suggests an ethics system for AI in society. This is only one example of the many aspects that require more clarity and greater focus when addressing the "Ethics-in-AI gap." A clear term of reference may also be missing.

1.2 Purpose and Objectives

The purpose of this White Paper is to explore foundations for global policy to address the Ethics-in-AI gap. It stresses the importance of reviewing the “state of the art,” broadly defined, and proposes a work plan designed to meet two overarching objectives, namely, to:

a. Articulate, integrate and frame core principles for a system of Ethics-in-AI sensitive to complexities of innovation and competition, as well as one responsive to the diversity of perspectives in the international community, and to the extent possible,
b. Address operational implications and challenges of Ethics-in-AI in diverse societal and technological contexts.

Each is considerably more complex than appears at first glance; and each anticipates the conduct of preliminary research and reviews of existing materials (see Section III on Program Design). Central to this White Paper is the importance of context.

1.3 Context and Logic

The basic proposition underlying this White Paper is that the parameters for agreement on ethics in artificial intelligence remain largely uncharted and fraught with diverse types of unknowns. It is developed in the context of an international collaborative initiative designed to frame and establish an Artificial Intelligence International Accord (AIIA). The initial framing was prepared for the United Nations centennial book Remaking the World Toward the Age of Global Enlightenment (2021) Chapter 2.

The leader in this initiative is the Boston Global Forum and the Club of Madrid (CdM) association of former Presidents of Democracies, and the United Nations Academic Impact. Jointly, they span communities of research, policy, and practice. The logic is this: the AIIA initiative has already attracted considerable interest and support from the policy, academic, and business communities in different parts of the world. The current thinking, however, does not incorporate coverage of ethics nor does it assume that ethics must be a central feature of any emergent international agreement on AI.

At the same time, the participants in the initiative recognize the ethics-gap in AI and appreciate that it must be addressed in an overall framing of the emergent Artificial Intelligence International Accord. This situation provides an important pragmatic and operational venue within which to extend beyond the many efforts made so far in the articulation of broad-based Ethics-in-AI.

1.4 Premises and Practice

This White Paper starts with a pragmatic stance—the world as we know it—captured by this set of premises:

- Continued innovations in science and technology
- Absence of constraints on research and exploring the "unknown" (subject to current legal, human, and social formal regulations)
- Sustained research on privacy and human rights
- Close review and assessment of machine-brain interactions (or interface systems)
- Respect for patents, copyright, and other supports for knowledge protection
- Attention to culturally-based ethical considerations

In practice, these provide a basis for situating aspirational objectives as, and if, they emerge. We anticipate these premises and their implications—as well as the overall project design (see Section III)—to be subject to appropriate external review and independent evaluation by:

- Establishing an independent advisory and review board
- Pursuing the usual academic and research processes related to publication
- Submitting research and results for peer review in the academic, business, and policy domain(s).

1.5 What Lies Ahead

The White Paper is organized as follows:

Section 2 introduces the defining imperatives, that is, the general principles of research governing the Project Design in Section III.

Section 3 presents the project design, including (i) more detailed statement of the objectives introduced in Section 1.3, and (ii) proposed logic and methods, with forms of computation and analytics.

Section 4 outlines the expected outcomes—products and processes—as well as overall value added.

Appendix introduces the collaborating entities.

2 Imperatives for Ethics in AI

To reduce the dangers of undue simplification, or the trap of "one size fits all," and to avoid implicit bias, this White Paper highlights four distinct, but interconnected, imperatives. They are a "basic checklist" to ensure that we remain on course, that is, working towards an integrated and coherent system of "Ethics-in-AI." They serve largely as important points of departure for reflecting on the fundamentals at hand.

These consist of: (i) Dimensions of Analysis, (ii) Domains of Interaction, (iii) Levels of Analysis, and (iv) Fundamentals in Foundations. Designed to provide methodological guidance, they ensure consistency and facilitate data organization and analysis, as well as the reporting of results. Jointly, they will help frame our understanding of coherence in "Ethics-in-AI" and may even provide a plausible causal logic with attendant dynamics. Now, we proceed to a brief note on each:

2.1 Dimensions of Analysis

The first imperative, dimensions of analysis, is threefold:

a. Conceptual and theoretical
b. Computational and analytical
c. Empirical and operational

The conceptual and theoretical dimension highlights characteristic features of the fundamental ethical issues, as understood in all social contexts. There might be disagreements across countries and cultures, but convergence on the basics may be more likely than unlikely.

Less clarity and even less consensus may emerge around the second dimension, computational and analytical. For example, transparency is generally considered a "good thing," and is often tied into scientific research and applications of technology. In market or other competitive contexts, transparency may become contentious, as trade-offs with other values become apparent. The third dimension, empirical and operational, may well be, at first glance, the least contentious in articulation or measurement.

The value of the above lies mainly in (a) identifying and capturing the diversity of ethical discourse in AI, (b) providing some coherence within prevailing discourse, and

(c) creating robust foundations for articulating coherence in the overall initiative.

2.2 Domains of Interaction

The second imperative is about the distinct domains of interaction, shaped by, and shaping, prevailing norms and behaviors everywhere. These are the overarching "spaces" within which humans interact:

a. Human Society – and its geopolitical considerations
b. Natural Environment – and its life supporting properties
c. Cyberspace – and its enabling and generative communication capabilities.

It cannot be assumed that innovations and applications of AI are contained within any one of these "spaces" – given that they are highly permeable and interconnected.

Nonetheless, to the extent possible, it is essential to recognize and "keep track" of sources and consequences of AI innovations and applications.

2.3 Levels of Analysis

The third imperative is to address the levels of analysis and the dilemmas created by the near ubiquity of AI applications in various scales and scopes. The very pervasiveness of AI may itself enable or promote the "one size fits all" bias. In this connection, four levels of analysis—common in the study of international relations—can be of considerable relevance:

- The Individual
- State and society
- International System
- Global system

While stylistically framed, these notions are empirically robust in articulating types of human aggregation.

The value of levels of analysis lies in highlighting (i) characteristic features of each aggregation, (ii) linkages connecting them and, with the exception of the global system, the (iii) scope of decision-making. Each of the first three levels includes systems and/or decision organizations, but there is as of yet no all-encompassing decision-system at the global level.

This exception may also be relevant as these levels harbor different sources of AI innovations and attendant applications.

2.4 Foundational Features

The fourth imperative draws attention to select foundational works essential for articulating "Ethics-in-AI" – over and above the usual literature reviews. These include:

i. History of AI—ongoing initiative of AIWS
ii. Atlas of AI—the "footprints" of AI
iii. AI in the Wild—environmental issues
iv. Artificial Intelligent You—relevance to the Individual
v. Thinking Fast, Thinking Slow—cognitive modes and models

These are important because they highlight different perspectives, on the one hand, as well as different facets of AI applications not generally addressed, on the other.

The value added is created by the distinctive light these works cast on the matter of AI ethics, even if ethics is not the direct focus. Note, for example, Atlas of AI provides important evidence about the impact of AI on the natural environment. This is most surely an important aspect of any consideration of ethics.

3 Program Design: Objectives, Methods, Value Added

The Program Design is structured around five objectives. What follows is an introduction to each objective, including methods used, expected results, and a statement of value.

3.1 One: Review of AI General "State of the Art"

Current understandings of AI are embedded in a wide but disparate range of literature, mainly of a technical nature (computational and related issues), with some contributions from the social sciences (cultural, social and normative issues). While most of the literature emanates from industrial societies, there is a notable growth of attention in developing countries.

The first objective is to (a) create some order in prevailing views of ethics-in-A, and (b) construct a coherent view of the current "state of the art" in the broad arena of AI. Tasks include:

- Reviewing existing frameworks for AI put forth at the national level in different parts of the world.
- Identifying existing methods to prevent abuses in uses of AI, data, digital technology, and the domain of cyberspace (including attacking companies, organizations, countries, and individuals on the Internet).
- Consolidating data on new norms to manage known aspects of AI innovations and applications.
- Exploring potential AI "unknowns" in order to anticipate the attendant or potential companion ethics.
- Reviewing the provisions of the Budapest Convention on Cybercrime as well as the EU General Directives, and/or incorporating basic principles thereof.
- Envisioning a sanctions system for violations of rights and responsibilities associated with development, design, applications, or implementation of AI.
- Mapping expressions of attention to, and relevance of, ethics statements as related to existing international agreements on human rights.

The expected result is a comprehensive and "best review" of the literature related to ethics in AI. The value-added lies in creating a "system boundary" for the substantive inquiry and issues raised in the overall program.

Further value is found in helping (i) steer global conversations on fundamental rights and responsibilities in digital societies, (ii) build bridges between countries, regions, and communities on matters of AI norms and ethics, and (iii) identify paths to consensus for a rights-based agenda of governance for AI and digital technologies.

3.2 Two: Situating Ethics

The second core program objective is to identify, as clearly as possible, the role of ethics as understood so far. This is done in two steps: First, locate where ethics are addressed in each of the tasks in Objective One. Second, map ethics-statements onto each of the four Imperatives introduced in Section II above.

When in doubt, we propose to identify the “counter-ethics,” that is, to what extent can we note the potential “bad,” i.e., what cannot or must not be done. This might appear simplistic, but even a cursory look at the literature illustrates that which “must not be done.”

The value of this work lies in situating the context of stated ethics and identifying content – aspirational as well as operational. The results of Objective Two are to be reported in the form of a database organized by key variables and sources (see below for expected uses).

3.3 Three: Policy Analytics for Ethics-in-AI

Objective Three focuses on policy responses—private and public, national and international—to AI challenges and matters of ethics. The term “policy” is used in its inclusive form here to include directives, guidelines, and legal statements. By definition, these are written in linear sequential text form—word after word, sentence after sentence, and so on—which makes them difficult to integrate and conceals policy- technology-security interactions.

The text form obscures embedded feedback, delays, or potential unintended consequences in the directives for action. Further, as concluded by RAND on a related issue years ago, there is dilemma: a large body of policy directives and documents have come with major barriers to understanding, deployment, and implementation. This dilemma is pervasive throughout the entire ecosystem of discourse on Artificial Intelligence directives and guidelines for ethics.

We propose to leverage well-known engineering tools—usually applied to operational and technical challenges—to transform text into data for metrics, thus transcending the constraints of text form. The process begins with (1) converting text into a Design Structure Matrix (DSM), then (2) creating a matrix of metrics as the reference model for network analysis, to explore and visualize connections among components. Network models with visualization tools represents content empirically. The full DSM for each text allows us to focus on segments thereof or “dig in deeper,” as needed, and statistical tools allow us to examine parts and pieces of the DSM.

The results of Objective Three include an empirical database of structure and substantive content, as well as the identification of embedded features for the entire text-corpus examined. The added value lies in aggregating results across texts to identify central tendencies, outliers, and other fundamental features of the database.

3.4 Four: Ontology of Ethics-in-AI

The fourth objective is to manage the inquiry and analysis of data on Ethics-in AI in ways that are consistent with the Imperatives, in Section II above. To the extent feasible, this means situating and recording data entries according to: (a) dimensions, (b) domains, and (c) levels of analysis.

The result of Objective Four is an empirically structured ontology of Ethics-on-AI that is (a) derived from the database and (b) will assist in organizing the database. For example, the database would be coded by (i) System Basics; (ii) System Ethics-related Challenges (and problems); (iii) Proposed Solutions; and (iv) Expected System Adjustments. A range of other options may well arise as central elements for coding purposes. The value added is that the ontology would also form the basis for a searchable knowledge repository on Ethics-in-AI.

3.5 Five: Ethics-in-AI and National Profiles

Objectives One to Four are centered on Ethics-in-AI. By contrast, Objective Five seeks to situate ethics within the national context. At this point, the White Paper raises the following question: Is there a relationship between the characteristic features of a country and its AI policies, including ethics? This question is addressed computationally (and statistically) by comparing (a) results of Objectives One to Four on ethics, to (b) data on activities of states in both the cyber and "real" domains. The value added lies in identifying state and corporate propensities, if any, for particular Ethics-in-AI configurations.

4 Expected Outcomes

The expected outcomes of the initiative on Ethics-in-AI, as presently conceived, are framed around three themes:

- First are results of the five objectives in Section III and the added-value of each activity.
- Second is an institutional outcome, envisioned in the form of a distributed AI Global Policy Lab.
- Third is a community development process addressing the future of Artificial Intelligence and Policy Analytics.

We now turn to each theme.

4.1 Anticipated Benefits of Five Objectives

Briefly stated, it is expected that:

- Objective One will yield a comprehensive and "best review" of the literature related to Ethics-in-AI. The value-added lies in creating a "system boundary" for the substantive inquiry and issues raised in the program design as a whole.
- Objective Two will identify the content of stated Ethics-in-AI, as well as the context. The results are to be reported in the form of a database organized by key variables and sources (see below for use). The value added lies in distinguishing between aspirational and operational postures on ethics.
- Objective Three is expected to generate an empirical database of AI policy structure and substantive content, as well as embedded features therein. The value added lies in aggregating results across texts to identify central tendencies, outliers, and other fundamental features of the database.
- Objective Four will yield an empirically-based structured ontology of Ethics-in-AI that is (a) derived from the database and (b) will assist in organizing the database. The value added is that the ontology would also form the basis for a searchable knowledge repository on Ethics-in-AI.
- Objective Five, will identify the connections (if any) between AI policies and state profiles. The results for Objective Five are to be reported in statistical terms. The value added lies in locating potential national propensities for particular Ethics-in- AI configurations. The term "national" in this context includes the corporate and business sectors.

4.2 AI Global Policy Lab

We expect this program to become the foundation for, and core of, the AI Global Policy Lab at MIT, in collaboration with the participating entities of this proposal. The purpose of the AI Global Policy Lab is to:

a. Address the development of efficient, cost-effective, accessible, and sustainable technical solutions, policy responses and ground support for challenges related to "Ethics-in-AI,"
b. Serve as a platform for delivering outreach programs on digital safety, responsible AI, and policy advocacy for "Ethics-in-AI," and
c. Enable students and others to explore alternative "pathways" and methods to increase their knowledge of, and sensitivity to, implications of innovation in AI.

4.3 Artificial Intelligence & Policy Analytics

The community development mission of the project aims to develop the next generation of researchers in the field of Artificial Intelligence & Policy Analytics. We will pursue a core set of programs that provide professional development opportunities, as well as resources and learning materials, for faculty and researchers w as well as policy analysts working in related fields.

We expect to co-organize (with the collaborators and supporting organizations) an annual meeting/technical workshop(s) that combine general reflections on ethics in the digital era with more issue-based policy discussions regarding specific areas where AI and digital technologies are driving profound societal transformation. These event(s) will be attended by researchers from MIT, collaborating entities, government representatives, and invited guests from the industry – to help represent and articulate diverse perspective on ethics in AI.

5 APPENDIX COLLABORATING ENTITIES

This Appendix introduces the lead collaborators in the proposed Ethics-in-AI Initiative. Most notable are the following:

5.1 Boston Global Forum (BGF)

BGF was founded to bring together thought leaders and experts from around the globe to participate in open public discussions and focus on the most critical issues affecting the world at large. Its main mission is to provide an interactive and collaborative forum for identifying and developing approaches to our most profound problems.

5.2 Club de Madrid (CdM)

An association of over 110 democratic former Presidents and Prime Ministers, CdM serves as a venue for exchanges on fundamental rights in digital societies. The purpose is to strengthen connections among countries, regions, and communities of practice, to work toward rights-based governance of AI and digital technologies.

5.3 United Nations Academic Impact (UNAI)

This initiative of the United Nations seeks to align institutions of higher education with the UN to support and contribute to UN goals and mandates, including human rights, education, sustainability and conflict resolution.