OPERATIONALIZING AI ETHICS PRINCIPLES *

Cansu Canca, Ph.D. †

AI Ethics Lab

Artificial intelligence (AI) has become a part of our everyday lives from healthcare to law enforcement. AI-related ethical challenges have grown apace ranging from algorithmic bias and data privacy to transparency and accountability. As a direct reaction to these growing ethical concerns, organizations have been publishing their AI principles for ethical practice (over 100 sets and increasing). However, the multiplication of these mostly vaguely formulated principles has not proven to be helpful in guiding practice. Only by operationalizing AI principles for ethical practice can we help computer scientists, developers, and designers to (1) spot and think through ethical issues; and (2) recognize when a complex ethical issue requires in-depth expert analysis. These operationalized AI principles for ethical practice will also help organizations confront unavoidable value trade-offs and consciously set priorities. At the outset, it should be recognized that by their nature, AI ethics principles—as any principle-based framework—are not complete systems for ethical decision-making and not suitable for solving complex ethical problems. But once operationalized, they provide a valuable tool for detecting, conceptualizing, and devising solutions for ethical issues.

With the aim of operationalizing AI principles and guiding ethical practice, we created the *Dynamics of AI Principles*,[i] an interactive toolbox with features to (1) sort, locate, and visualize sets of AI principles demonstrating their chronological, regional, and organizational development; (2) compare key points of different sets of principles; (3) show distribution of core principles; and (4) systematize the relation between principles.[ii] By collecting, sorting, and comparing different sets of AI principles, we

---

* Written and submitted in Spring 2020.
A version of this material will appear in the December *Communications of the ACM* Computing Ethics column (volume 63, number 12).
† Founder and Director of AI Ethics Lab; aiethicslab.com/cansu-canca/; cansu@aiethicslab.com

discovered a barrier for operationalization: many of the sets of AI principles mix together *core* and *instrumental* principles without regard for how they relate to each other.

In any given set of AI principles, one finds a wide range of concepts like privacy, transparency, fairness, and autonomy. Such a list mixes core principles that have intrinsic values with instrumental principles whose function is to protect these intrinsic values.[iii] Human autonomy, for example, is an intrinsic value; it is valuable for its own sake. Consent, privacy, and transparency, on the other hand, are instrumental: we value them to the extent they protect autonomy and other intrinsic values. Understanding these categories and their relation to each other is the key to operationalizing AI principles that can inform both developers and organizations.

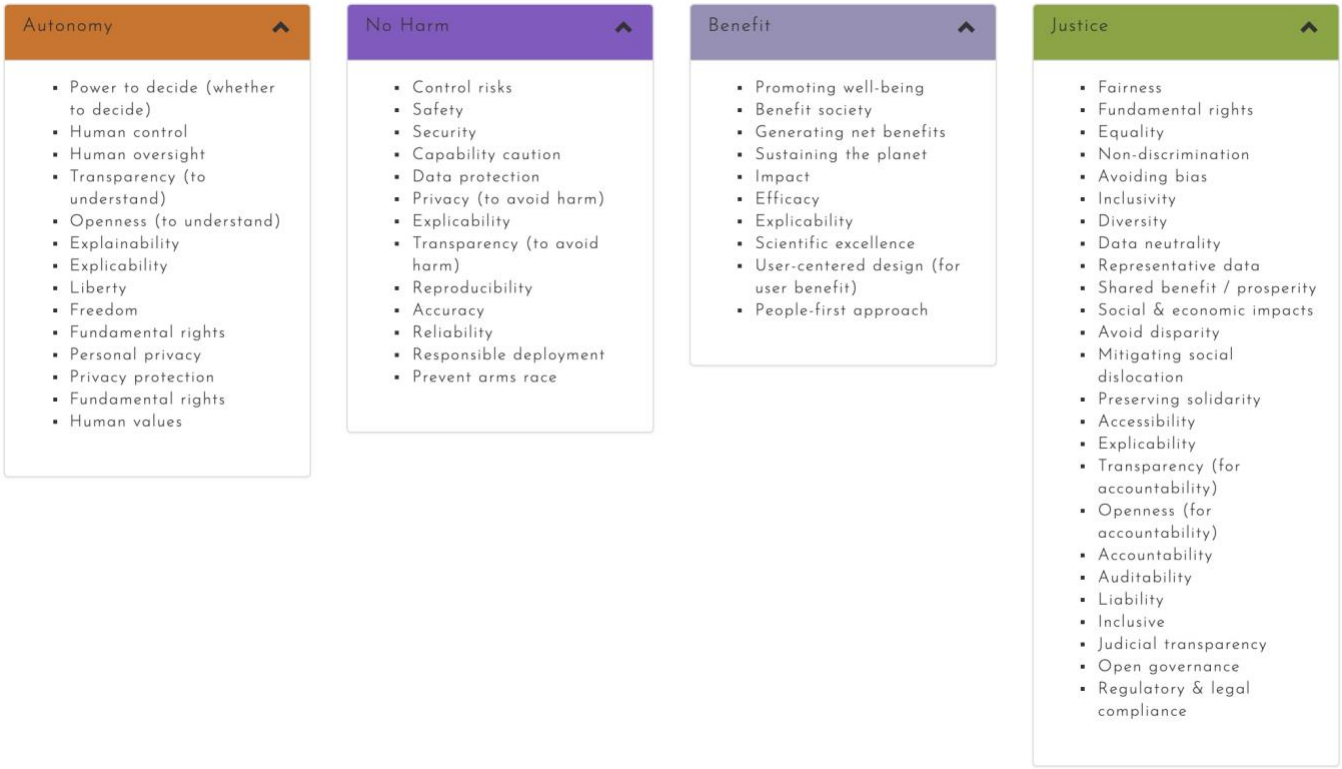
Core versus Instrumental Principles

The most widely utilized set of core principles in applied ethics is: respect for autonomy, beneficence (avoiding harm and doing good), and justice.[iv] These principles prescribe an appropriate attitude towards certain values: *respect* autonomy, *do* good, *ensure* justice. They are *core* principles because they invoke those values that theories in moral and political philosophy argue to be *intrinsically* valuable, meaning their value is not derived from something else. These theories answer the central question of applied ethics: "what is the right/good action or policy to choose?" By encapsulating these theories' intrinsic values in an attitude-setting format, core principles help us immediately recognize if we are facing an ethical challenge: Any action that disrespects autonomy, inflicts harm, or discriminates is ethically problematic, if not straightforward unethical.

When we categorized all of the published AI principles into these three core principles, we found a surprisingly balanced and consistent picture. Across industries and regions, similar weight is given to each of these principles and neither one really outweighs the others: About 25–30% of all principles are autonomy-focused, about another 32–34% are focused on avoidance of harm and maximizing benefits, and about 36–41% of principles are justice-focused.[v]

In contrast to core principles encapsulating intrinsic values, instrumental principles build on concepts whose values are derived from their instrumental effect in protecting and promoting intrinsic values. Take transparency for example. We do not think that something is valuable in and of itself solely

because it is transparent. Rather transparency is valuable because it allows us to understand, engage with, and audit the systems that affect us. In other words, transparency is instrumental to uphold intrinsic values of human autonomy and justice. Similarly, accountability in itself is not an end but rather it is a means to safeguard justice by assigning responsibility and to avoid harm through deterrence.



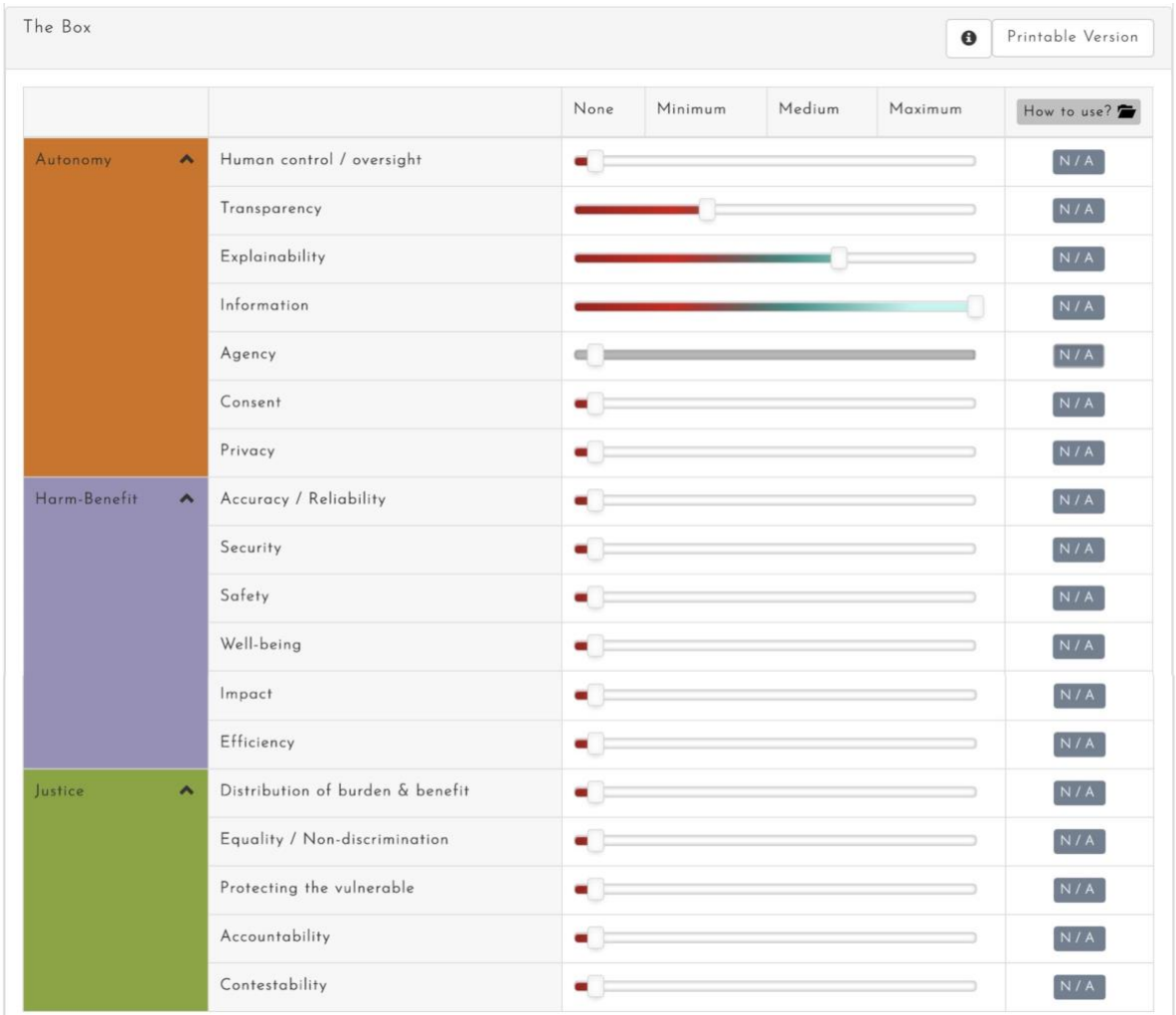
Operationalizing the AI Principles to Guide AI Practice

To guide AI practice, it is important to distinguish core and instrumental principles because instrumental principles are interchangeable. Individuals and organizations can weigh instrumental principles to determine which to prioritize and how to use them to best achieve the core principles.

For example, the instrumental principle of explainability may be required or optional for ethical AI depending on the situation. Explainability is about understanding a system's technical process and how it reaches its outcome. It is an instrumental principle that can uphold human autonomy by allowing individuals to interact meaningfully with the system. It can also help minimize harm and safeguard justice by making it easier to detect errors and biases. Explainability is crucial for a risk analysis AI system that helps judges set bail and parole in the criminal justice system. Here

explainability enables judges and defendants to engage autonomously with the system and better monitor its fairness. In contrast, explainability is not necessary for an AI system that optimizes energy use in cars. Drivers need outcome- and safety-related information to exercise their autonomy, and developers need rigorous testing for safety and accuracy to minimize harm. Explainability in this case should not be prioritized especially if it would compromise accuracy and safety.

The general point is that there is no deep ethical dilemma when an instrumental principle is not suitable for a given case. Once we correctly categorize principles as core and instrumental, we can turn many vague AI principles into an operational checklist to guide practice. We created a simplified checklist called *the Box* for computer scientists, developers, and designers to use for basic ethics analyses. The Box helps them determine the relevant ethical concerns and weigh applicable instrumental principles to determine how to best satisfy core principles by substituting or supporting one instrumental principle with another.



The Box also serves as a tool for computer scientists, developers, and designers to recognize when an ethical conflict is between core principles. Since each core principle is intrinsically valuable, we cannot simply ignore one for the sake of another. When core principles point in opposite directions, we face a real ethical dilemma. In these cases, an ethical issue is complex and cannot be easily resolved solely based on guidance from principles; ethics expert should be brought in to apply ethical theories. An example could be developing an AI system for diagnosing a dangerous and highly contagious disease. To minimize harm and suffering, scientists need to quickly create a highly accurate system. To train the algorithm they need large amounts of personal data that cannot be completely anonymized. Asking for proper consent would delay the project causing more infections and more suffering. However, circumventing consent and disregarding privacy would be a violation of individual autonomy. An AI ethics framework that relies solely on principles is unable to solve this ethical dilemma suggesting that this complex problem should involve ethics experts who can conduct an in-depth analysis and apply ethical theories. Principle-based frameworks only provide a list of considerations rather than a complete and coherent decision-making tool.[vi] Ethical theories, on the other hand, provide comprehensive guidance for decision-making and ethics experts can utilize these detailed theories to analyze complex issues in-depth.

When properly operationalized, AI principles provide a helpful start for an ethics analysis and they can guide developers and organizations through many ethical questions, even though they are not sufficient for complex ethical problems. We need to understand AI principles for what they are: A list of fundamentally and instrumentally important ethical considerations but *not* a complete system for complex ethical decision-making. Categorizing and using AI principles ensures that we do not overlook a crucial ethical concern. By revealing when a case presents a conflict between core principles rather than instrumental ones, principles can also help us recognize when we face a complex case and need a full-scale ethics analysis. As we operationalize AI principles, we need to utilize their strengths and recognize their limitations, acknowledging that AI principles are only a first step for development and deployment of ethical AI.

Setting an Organization's AI Principles

Lastly, let us go back to the proliferation of AI principles. What is the point of company-specific sets of principles if the content is largely the same, as we have seen? If well-done, the point is that

organizations with their own customized set of AI principles can determine how to weigh competing principles and which intrinsic value to prioritize when core principles (and theories) conflict. When no single argument emerges as the strongest even after a full-scale ethics analysis, developers and other organizational decision-makers must choose between equally ethically permissible options. This is when the organization's stance about core principles shows through. When push comes to shove, does this organization prioritize individual autonomy or minimization of harm? An organization's AI principles can guide this decision and be useful both for computer scientists and ethics experts when they are deciding a hard case. To help them clarify their own values, we invite organizations to use our toolbox to compare their efforts to others, systematize their AI principles, and engage in in-depth ethics analysis with experts to determine their own priorities for coherent and consistent ethical decision-making.

FOOTNOTES

[i] AI Ethics Lab, *Dynamics of AI Principles*, published in February 2020, http://aiethicslab.com/big-picture/.

[ii] Similar efforts have been done before and we have cross-checked our list with these other works. See, Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI", *Berkman Klein Center Research Publication* 1 (2020); Jobin, Ienca, & Vayena, "The Global Landscape of AI Ethics Guidelines", *Nature Machine Intelligence* 1, 389–399 (2019); Zeng, Lu, Huangfu, "Linking Artificial Intelligence Principles", *AAAI Workshop on Artificial Intelligence Safety* (2019). These works look at the frequency of principles rather than categorizing them conceptually.

[iii] The only document that conceptually categorizes principles is the paper by AI4People, which recites the four "core" principles (beneficence, non-maleficence, autonomy, and justice) and adds the principle of explicability as an "enabling" principle. Floridi et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", *Minds & Machines* 28, 689–707 (2018).

[iv] In 1978, United States National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research published the *Belmont Report*. The *Report* laid out three core ethical principles for human subject research: respect for persons, beneficence (which divides further into two general rules of "do not harm" and "maximize possible benefits and minimize possible harms"), and justice. In 1979, philosophers Tom Beauchamp (who was at the Commission co-authoring the *Report*) and James Childress published the canonical book *Principles of Biomedical Ethics*, where they identified four prima facie ethical principles: respect for autonomy, beneficence, non-maleficence, and justice (see, Beauchamp and Childress, *Principles of Biomedical Ethics*, 8th edition, Oxford University Press: 2019). Principles in the book and in the report overlap in their content and form what is now often called the "traditional bioethics principles."

[v] The principle of justice holds the largest pie (about 40%). This is not surprising because the justice principle is much more general than the other two principles referring to various theories of justice and their main fairness claims such as equal treatment, equal opportunity, and protection of the worst off.

[vi] An excellent paper on this is Clouser and Gert, "A Critique of Principlism", *Journal of Medicine and Philosophy*, 15(2): 219-236 (1990).