Peace
AI

Ethics in AI for Peace

Towards Ethical Design and Implementation of AI in Peacebuilding

AI ETHICS CARDS

AI ETHICS CARDS

Why Cards?

Cards are a concise and accessible way to present key ethical principles and questions. They serve as quick reminders and prompts to guide decision-making, making it easier for stakeholders to integrate ethical considerations into AI development and deployment.

A Framework of Principles and Reflective Questions for Ethical AI Design and Implementation

The AI Ethics Cards represent a core set of principles that AI for Peace is committed to upholding in the pursuit of responsible and ethical AI development.

Each card highlights a key ethical principle and is accompanied by a set of guiding questions designed to help developers, policymakers, and other stakeholders navigate complex ethical landscapes.

These questions encourage thoughtful reflection and practical application, ensuring that critical ethical considerations are addressed before the design, development, or deployment of AI systems.

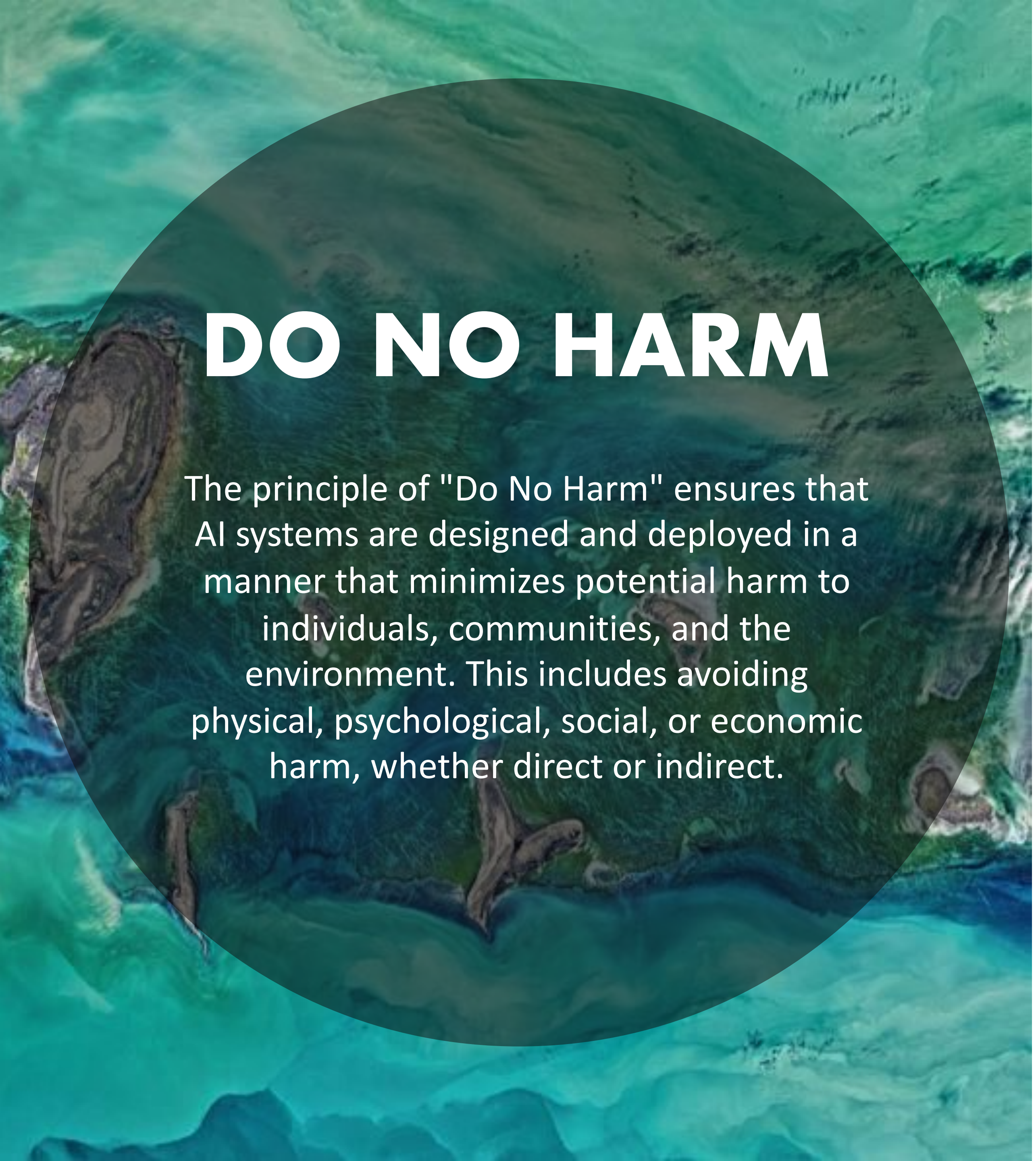
By integrating these principles, we aim to foster AI systems that promote peace, inclusivity, and the well-being of all communities.

ETHICS BY DESIGN

Ethics by design incorporates ethical considerations and principles from the earliest stages of the design process. The goal of ethics by design is to ensure that technology is developed in a way that is aligned with societal values and that minimizes negative impacts on individuals and communities.

Guiding questions:

- □ What **core values** will guide the design, development, and implementation of your AI solution?
- □ Is your team familiar with established **AI principles**, and have you selected a specific framework to adopt?
- □ What measures will you take to ensure these values and principles are effectively implemented?
- □ Will you utilize a specific ethics tool or framework to support this process?



DO NO HARM

The principle of "Do No Harm" ensures that AI systems are designed and deployed in a manner that minimizes potential harm to individuals, communities, and the environment. This includes avoiding physical, psychological, social, or economic harm, whether direct or indirect.

Guiding questions:

- □ Could the AI system intentionally or unintentionally harm affected groups?
- □ How does the system ensure resilience against adversarial attacks or misuse?
- □ What safeguards are in place to mitigate risks or unintended consequences?
- □ Have potential harms been assessed across short-term and long-term timelines?
- □ What measures are taken to ensure the system's misuse is prevented?

CONFLICT SENSITIVITY

The principle of "Conflict Sensitivity" ensures that AI systems are developed with an understanding of the contexts they operate in, particularly in regions with tensions or ongoing conflicts. The goal is to avoid exacerbating conflicts and to promote peaceful outcomes.

Guiding questions:

- ❑ How well does your team (and the system you are building) understand the context in which you operate?
- ❑ Does your team (and the system) understand the results of the interactions between your activities and that context?
- ❑ What steps are being taken to minimize the negative impacts of your work (and the final AI system's)?
- ❑ How can your team (and the system you are developing) maximize its positive effects for peace?

BIAS AND FAIRNESS

Addressing bias and fairness involves identifying and rectifying potential biases in training data and algorithms to ensure that the AI system's predictions and recommendations are equitable and do not disproportionately disadvantage specific groups, including those based on gender or identity.

Guiding questions:

- □ Who is involved in the design and testing process, and how diverse is this group?
- □ Which groups could be disproportionately affected by the decisions made by the AI system?
- □ What strategies will you use to detect and address biases in the training data and algorithms to prevent discrimination?
- □ What steps will you take to involve affected communities in identifying potential biases or fairness concerns?
- □ How will you evaluate and maintain fairness in the system’s predictions and recommendations?

PRIVACY AND DATA PROTECTION

Privacy and data protection encompass safeguarding user data, obtaining informed consent, and implementing measures to protect sensitive information, including gender-related details, to respect user privacy and prevent unauthorized access or misuse.

Guiding questions:

- □ What types of data will the AI system collect, process, and store, and is this data collection necessary for its purpose?
- □ Do users have control over their data, and how will you obtain informed consent for data usage?
- □ How will you protect sensitive information, particularly when it comes to gender, identity, and personal preferences?
- □ What measures are in place to secure the data from unauthorized access or breaches?
- □ Will the data be anonymized or pseudonymized to protect individual identities?

PARTICIPATION AND COLLABORATION

Promoting participation and collaboration entails involving diverse stakeholders, including affected person by your solution, in the development process to ensure diverse perspectives are considered.

Guiding questions:

□ Who are the key stakeholders, and how will they be involved in the design, development, and deployment of the AI system?

□ What actions will you take to ensure that marginalized or underrepresented groups have a voice in decision-making?

□ What mechanisms will you use to gather input from affected communities and integrate their perspectives into the process?

□ How will you ensure transparency and effective communication with stakeholders throughout the project lifecycle? What tools or platforms will you use?

□ What steps will you take to foster collaboration across disciplines, sectors, or regions for holistic solutions?

HUMAN-CENTERED DESIGN

Prioritizing human-centered design means focusing on user needs, well-being, and values throughout development, while considering potential impacts on user autonomy, dignity, and agency, including those related to gender and identity.

Guiding questions:

- □ How will you prioritize user needs and well-being throughout the development process?
- □ Are there potential risks to user autonomy, dignity, and agency, and how can they be mitigated?
- □ How can the AI system contribute positively to users' lives and respect their values, including gender-related concerns?

TRANSPARENCY AND EXPLAINABILITY

Ensuring transparency and explainability means making the decision-making process of the AI system understandable and providing clear explanations for its conclusions or recommendations, especially in complex models, to enhance user trust and accountability.

Guiding questions:

□ What information about the AI system will you share with users, stakeholders, and the public?

□ How will you document and communicate the system's design, functionality, and decision-making processes?

□ What steps will you take to ensure the system’s limitations, uncertainties, and potential risks are clearly communicated?

□ What tools or methods will you use to improve explainability, such as model interpretability frameworks?

ACCOUNTABILITY AND RESPONSIBILITY

Establishing accountability and responsibility involves defining who is responsible for the AI system's outcomes, addressing errors or biases that arise, and creating a framework that ensures ethical decision-making throughout the project's lifecycle.

Peace AI

Guiding questions:

- ❑ Who is accountable for the AI system's decisions, outcomes, and impacts, and how will this be communicated?
- ❑ How will you document and track decisions made during the AI system’s lifecycle, including its design, development, and deployment?
- ❑ What actions will you take to ensure team members understand their ethical and legal responsibilities?
- ❑ Who will oversee the implementation of ethical guidelines and ensure they are adhered to?
- ❑ What mechanisms will you use to learn from mistakes and integrate lessons into future development?

Peace

AI