Global
Future
Councils

WORLD
ECONOMIC
FORUM

AI Value Alignment:
Guiding Artificial
Intelligence Towards
Shared Human Goals

WHITE PAPER

OCTOBER 2024

Contents

Foreword

Virginia Dignum
Professor of Responsible Artificial Intelligence, Umeå University; Co-Chair, Global Future Council on the Future of AI, World Economic Forum

As artificial intelligence (AI) systems become more integrated into various aspects of society, ensuring that they align with human values is critical. This paper explores AI value alignment, emphasizing the integration of ethical principles such as justice, privacy and fairness into AI technologies. We present practical frameworks and methodologies to ensure that AI systems uphold these values throughout their life cycle. By promoting a collaborative, transparent approach, we aim to guide a development of AI that is both innovative and ethically sound.

Despite significant advances in AI technologies, the concept of value alignment – ensuring that AI systems behave consistently with human values and ethical principles – requires greater understanding. This white paper attempts to address this critical issue, underscoring the need to further clarify many aspects of value alignment and to develop clear and standardized approaches. As AI continues to permeate various sectors, establishing comprehensive frameworks and guidelines is essential to ensure that these systems operate within acceptable ethical and societal norms. Standardization will not only facilitate transparency and accountability but also foster trust among stakeholders, thereby promoting the responsible and ethical deployment of AI technologies.

The Global Future Council on the Future of AI seeks to address and contribute to this important conversation, while highlighting concrete ways to address value alignment and promote a transparent and collaborative approach to AI development, use and governance in the future.

Executive summary

The concept of AI value alignment is essential to ensure that AI systems behave in ways consistent with human values, ethical principles and societal norms.

This white paper covers the concept of value alignment, including its definition and practical application and the processes involved in embedding values into artificial intelligence (AI) systems. Human values such as justice, privacy and agency are contrasted with such operational attributes as robustness and transparency, highlighting the importance of balancing ethical implications with technical mechanisms.

Exploring the entire life cycle of AI systems, the paper's analysis emphasizes the need for explicit and auditable processes to translate values into norms and verify their adherence. Active stakeholder participation and continuous monitoring are crucial to maintain alignment with societal values and ethical standards. A comprehensive

approach to AI value alignment also includes a detailed examination of frameworks, guidelines, human engagement, organizational change and auditing processes. These enablers help to ensure that AI systems are not only innovative but also ethical and trustworthy, thereby promoting trust and transparency among users and other stakeholders.

Finally, value alignment is linked to the concept of AI ethics red lines – the non-negotiable boundaries that AI systems must not cross. By embedding core human values and maintaining rigorous oversight, the value alignment process makes sure AI systems operate within established moral and legal frameworks, safeguarding against unethical behaviour and maintaining societal trust.

Introduction

Collective action on value alignment is crucial to ensure that AI systems reflect fundamental human values.

Human values such as justice, privacy and agency are fundamental principles that underpin ethical and moral frameworks in society. They are essential to protect human dignity and individual rights and to promote equitable and autonomous interactions. In contrast, operational attributes such as robustness, transparency and explainability, often included as non-functional requirements for systems, define the quality and performance of technological solutions. While crucial for building trustworthy and reliable AI systems, they are more about the system's operation and reliability than the ethical implications of its impact on human lives. Guidelines for responsible and trustworthy AI tend to blend these distinct types of values, which sometimes blur the line between human-centred ethics and the technical aspects of system design and implementation.

The concept of value alignment has emerged as a critical area of focus in AI. This concept revolves around making sure that the behaviours, decisions and outcomes of AI systems are in harmony with human values, ethical principles, societal norms and fundamental human rights.

Value alignment is fundamentally about human accountability, emphasizing that humans remain responsible for the ethical and societal impacts of AI systems, even as these systems become more intelligent than humans and control becomes an issue. While computational implementations can support and facilitate this alignment by embedding core values such as fairness, transparency and privacy into the system, they do not absolve humans from their ultimate responsibility. The process involves continuous monitoring, stakeholder involvement and compliance audits to ensure that AI systems adhere to established ethical standards and societal norms.

Mechanisms must be in place to maintain human oversight and decision-making authority. This includes implementing fail-safes, creating transparent decision-making processes and establishing clear protocols for human intervention when necessary. The importance of these measures is heightened in scenarios where AI could make autonomous decisions with significant ethical implications.

This white paper explores the fundamental aspects of value alignment, providing an in-depth look at the processes involved. First, a value taxonomy – showing different types of human values and preferences with illustrative examples in healthcare, credit scoring and autonomous driving – is discussed. This is followed by an examination of how value alignment can be applied in practice, exploring its significance at different stages of AI system development in different contexts and within various communities. The investigation covers essential tools for value alignment, such as frameworks, guidelines and methodologies, which are crucial for ensuring that AI systems align with human values.

Finally, the paper considers the critical link between value alignment and AI red lines in responsible AI development. Red lines represent the non-negotiable boundaries or ethical limits that AI systems must not cross. Understanding the relationship between value alignment and red lines is essential to comprehend the full scope and significance of responsible AI development and its impact on society.

Through this exploration, the white paper provides an overview of the current state of value alignment, offering insights into its importance, application and implications for the future of technology and society.

① What is AI value alignment?

Defining value-aligned AI systems involves several dimensions, including ethical principles, cultural contexts and societal impact.

AI value alignment is crucial for ensuring that the behaviour of AI systems is consistent with human values. Human values are linked to fundamental principles that guide human behaviour and decision-making and may encompass concepts such as justice, privacy, autonomy and respect, serving as the ethical foundation on which societies are built.

Universal human values tend to be highly abstract and can be interpreted differently across cultures, communities and situations. For example, while respect is a universal value, its interpretation can vary greatly, from expecting a handshake to expecting no physical contact at all. Existing value systems tend to vary in their approach to defining human values. A value taxonomy (outlined in section 2.1.1) can help categorize human values and preferences, illustrating how they can be understood and implemented in various contexts.

AI value alignment requires that the entire process – from translating values into norms, implementing these norms and verifying their adherence – is explicit and auditable. This means that every step must be clearly documented and open to scrutiny so that transparency and accountability can be checked.

Effective value alignment also requires active participation from different stakeholders so that the necessary interpretations of values and the outcomes of these interpretations are properly understood. This participatory approach ensures that the AI system aligns with the values of the community it serves.

To build a value-aligned AI system, there are many dimensions to consider and address:

– What human values the system should be aligned with.

– What it means to be aligned with such values.

– How this alignment is to be achieved.

– How to verify that the system is indeed aligned.

– How to monitor possible drift in alignment over time after the system's deployment. How to update the system if the values change.

– How to update the system if the values change.

An AI system used in a hospital setting provides support on patient diagnosis and treatment recommendations. Ensuring value alignment in this context means that the AI system must uphold core human values such as patient autonomy, privacy, fairness and human agency. The human agency here is that of the doctor who will make the final decision based on the AI recommendation. Additionally, the system must address the asymmetry of information between patients and healthcare providers so that patients are fully informed about how the AI system operates and how its recommendations are generated, thereby promoting transparency and trust.

Core human values

1. **Patient autonomy:** The AI system should respect the decisions and preferences of patients. For instance, if a patient chooses a less aggressive treatment due to personal beliefs, the AI should support this decision by providing relevant information about the risks and benefits without coercion.

2. **Privacy:** The AI must protect patient confidentiality so that sensitive health information is securely stored and accessible only to authorized personnel. This aligns with the fundamental human right to privacy.

3. **Fairness:** Recommendations by the AI system should be unbiased and equitable. For example, it should not discriminate against patients based on race, gender, socioeconomic status or any other factors.

4. **Trust:** The AI system must be transparent, reliable and accountable in order to foster trust between healthcare providers and patients and to make sure there is confidence in its recommendations.

Other values
(non-functional requirements)

1. **Compliance with regulations:** The AI system must comply with healthcare regulations – for example, the Health Insurance Portability and Accountability Act (HIPAA) in the United States mandates strict standards for data protection and patient privacy.

2. **Technical robustness**: The system should be technically robust, meaning it must be reliable and accurate in its diagnoses and recommendations. This involves rigorous testing and validation to prevent errors that could harm patients.

3. **Interoperability:** The AI system should be able to integrate seamlessly with existing hospital systems, such as electronic health records (EHR), to provide comprehensive care without disrupting the workflow of healthcare providers.

Implementation and evaluation

The values of patient autonomy, privacy and fairness can be applied through decision-support tools, encryption and diverse training datasets, while compliance, robustness and interoperability can be ensured via regulatory checks, rigorous testing and standard data formats. Evaluation can be carried out through patient satisfaction surveys, security audits, bias metrics, external audits, performance monitoring and interoperability testing.

1.1 | Value alignment and cultural differences

Cultural differences can also affect value prioritization, with variations seen, for example, in how different societies prioritize individual privacy vs. collective harmony.[1] Furthermore, each person may have their own values and preferences, which may differ from those of others. Therefore, human values may be identified at the level of each individual, each organization, each nation or globally. Identifying and prioritizing the right values at various levels is essential, depending on the deployment scenario. This is challenging for AI systems deployed at one level, such as globally or within a nation, when the value alignment has been implemented at a different level. Examples are included in Figure 2.

FIGURE 2 | Cultural context and individual differences

To address the cultural context and individual differences in AI effectively, systems must be tailored to reflect diverse values and practices across various domains, as described in the following examples.

Healthcare

In healthcare, an AI system designed for patient diagnosis must prioritize patient autonomy, privacy and fairness. However, the interpretation of these values can differ greatly across cultures. In some cultures, patient autonomy often means providing individual patients with all the available information and allowing them to make their own healthcare decisions, while in other cultures it is common practice for family members to be involved in the decision-making process, reflecting a collective approach to patient autonomy. To address this, AI systems should incorporate decision-support tools that respect patient preferences and allow input from family members where appropriate. Evaluating the system's success should involve surveying patient satisfaction and measuring the system's ability to support culturally specific decision-making processes.

Credit scoring

Human values can mean different things in different situations. For example, to ensure fairness in an AI credit-scoring system, datasets that include diverse demographic groups should be used to train the model, taking into account cultural and individual differences in financial behaviour. Different cultures might have varying approaches to credit usage and savings, shaped by social, economic and historical factors, which should be reflected in the training data. Audits and fairness metrics – such as the disparate impact ratio[2] – should be conducted to evaluate the AI's performance across different demographic groups so that it does not disadvantage any particular group.

Autonomous driving

In the context of autonomous vehicles, ensuring that AI systems prioritize human safety involves implementing rigorous testing protocols and real-time monitoring to prevent accidents, with consideration for context-specific safety standards. For instance, traffic patterns and driving behaviours can vary significantly between urban and rural areas or between countries. Redundancy mechanisms and real-time monitoring systems can be used to enhance safety, adapting these measures to different driving environments and cultural norms regarding road safety, such as varying speed limits and pedestrian behaviours. System performance can be monitored through uptime metrics[3] and by conducting post-implementation reviews for safety incidents so that the system consistently meets safety expectations.

1.2 | Technical and organizational considerations in value alignment

Approaches to AI value alignment include both technical mechanisms and organizational processes/considerations.

In terms of organizational processes, value alignment usually begins with analysing deployment scenarios, engaging in multistakeholder consultations and training developers and users, among other factors.

Various technical approaches and mechanisms exist for checking and achieving value alignment. From the curation of training datasets to inverse reinforcement learning (IRL)[4] and reinforcement learning with human or AI feedback (RLHF),[5] developers have been building tools for value alignment for many years as AI techniques have evolved.[6] Such tools can be used in the various phases of AI development, including design, data collection, training, testing, deployment, use and audit.

Fairness is a value with significant implications for AI. Even before generative AI techniques were available, AI classifiers and predictors were scrutinized for fairness, with detection and
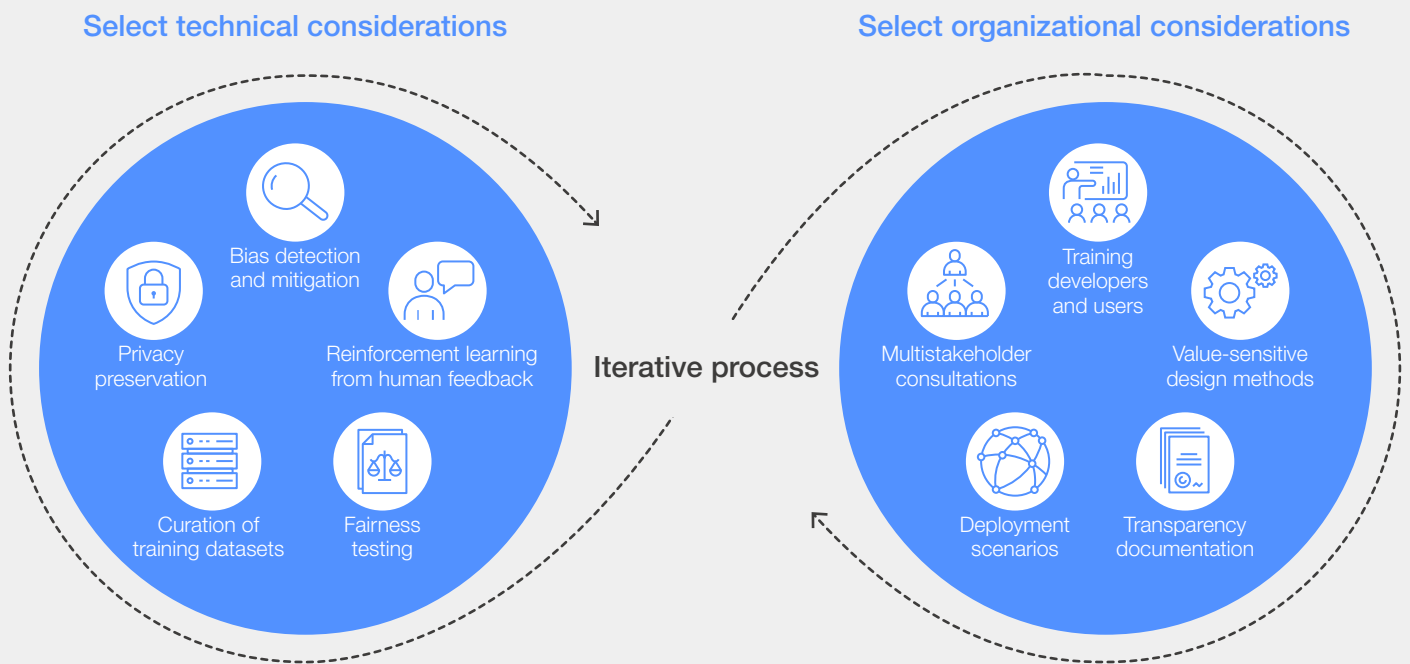
mitigation approaches in which the technology, AI itself, was used to identify and reduce bias. However, these technical solutions should be supported by multistakeholder consultations, developers' education, team diversity, community engagement, AI governance frameworks, research studies, sensitivity training and other organizational processes to ensure that the appropriate notion of fairness was selected for the specific decision scenario in which the AI system was intended for us. This long list of necessary activities shows that, even if the focus is on just one value, there is a great deal to study, discuss and decide at both societal and technical levels.

Societal agreement is necessary for validating AI value alignment. In the case of fairness – assuming the right notion of fairness for an AI system has been chosen – what threshold of bias is acceptable for certifying fairness alignment? In this regard, a closely related AI property is transparency: users of an AI system as well as regulators and auditors need to know what the developers did to check value alignment in the system. Without transparency, AI systems might be used that are not value-aligned at a level that is acceptable to users, or users might distrust AI systems that actually are sufficiently value-aligned because they have no way of knowing that.

The output of the value alignment process goes beyond a value aligned AI system to include transparency documentation covering which values have been considered, which value alignment tests have been carried out, which techniques have been used to achieve value alignment and which deployment and use scenarios are considered appropriate for the system. Importantly, this process is interdependent and iterative, constantly incorporating learnings from previous deployments to refine and improve the technical and organizational alignment of AI systems over time. Select technical and organizational considerations to achieve AI value alignment are detailed in Figure 3.

FIGURE 3 | **Select technical and organizational considerations in value alignment**

Select technical considerations

- Bias detection and mitigation
- Privacy preservation
- Reinforcement learning from human feedback
- Curation of training datasets
- Fairness testing

Iterative process

Select organizational considerations

- Training developers and users
- Multistakeholder consultations
- Value-sensitive design methods
- Deployment scenarios
- Transparency documentation

Source: World Economic Forum

2 | Value alignment in practice

AI value alignment involves continuous integration of human principles throughout the AI development life cycle.

2.1 | Context and communities

Shaping AI to align with human values transforms it from a tool for private interests into a technology that benefits humanity. This process, however, encompasses far more than merely adding community interests as an input into the AI alignment process as yet another checklist item. Rather, AI development should prioritize community interests, focusing on protection, identity preservation and practical solutions to real problems. Achieving this necessitates a comprehensive exploration that combines technical expertise with a human-centric perspective, paying special attention to existing and potential risks. This approach requires a strong commitment to understanding diverse human cultures and interdisciplinary collaboration. Community interests and values evolve, necessitating flexible and inclusive AI-alignment strategies that adapt over time with stakeholder input. A design philosophy centred on community interests enables AI systems to avoid harm and enhance community well-being. This approach allows for a broadening of vision and for collaboration across disciplines, ensuring that AI becomes a tool for enhancing collective well-being rather than merely avoiding pitfalls.

Value taxonomy

In order to ensure value alignment in AI systems, it is crucial to understand that human values are multifaceted, multicultural, multidisciplinary and context-dependent. There have been different approaches to defining human values, ranging from the highly aspirational Universal Declaration of Human Rights to the more specific laws and regulations of individual countries and jurisdictions, to common practices and social norms in different societies and to particular human preferences for each AI system application. The taxonomy of different value systems includes:

International instruments:

– The 30-point Universal Declaration of Human Rights has been accepted and signed by all member states of the United Nations (UN) and is widely regarded as the highest aspiration for all societies in guaranteeing the fundamental rights of each human.[7]

– The 17 Sustainable Development Goals (SDGs) of the UN are widely held objectives that can be referred to for building AI applications.[8]

– Lethal autonomous weapons (LAWS), like all other weapons, should be governed by the 1947 Geneva Conventions and their two 1977 Additional Protocols.[9] AI systems can be designed to align with these conventions and their underlying values.

– The Council of Europe adopted the Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, the first ever internationally legally binding treaty to promote human rights, democracy and the rule of law.[10]

– The Organisation for Economic Co-operation and Development (OECD) AI Principles are standards for AI based on fundamental human rights and democratic values.[11]

Ethical principles and moral philosophy approaches:

– Ethical principles and moral philosophy approaches, proposed and studied by philosophers for thousands of years, are the foundation not only of many accepted human values and social norms but also of hundreds of published AI ethics principles and guidelines. Western ethical principles are commonly derived from moral philosophy approaches such as utilitarianism, deontology or virtue ethics, among others,[12] and influenced by religious traditions such as Christianity, Islam and Judaism. Eastern philosophy and religious traditions – including Confucianism, Buddhism and Hinduism – can lead to a different emphasis in ethics;[13] African

traditions might lead to yet another set of ethical priorities.[14] The existing published AI ethical guidelines and regulations – such as the European Union (EU) AI Act,[15] the Chinese AI governance[16] position and guidelines[17] and the United States AI Bill of Rights[18] – are directly or indirectly based on the respective moral philosophy traditions of their place of composition. For the world to reach a consensus on AI governance, it is important for all countries to learn about and seek to understand other cultures' moral traditions.

National laws and regulations:

– AI systems need to be designed in accordance with the laws and regulations of the country in which they operate. In Thailand, the lèse-majesté law mandates that an AI system cannot produce content that would defame, insult or threaten the monarch of the country; in Singapore, the Protection from Online Falsehoods and Manipulation Act 2019[19] prohibits fake news or false information being created or spread by AI systems. Data protection laws also vary among jurisdictions, with the General Data Protection Regulation (GDPR)[20] in the EU, a patchwork of federal and state laws in the US,[21] the European AI Act[22] and so on.

Social norms and common practices:

– To gain user trust, AI systems need to be designed to comply with social norms and common practices, as long as they do not violate universal human values or the laws and regulations of the land. Social norms and common practices can include cultural

and social etiquette that both humans and AI systems are expected to follow, as well as behaviour that might be considered offensive in a particular society. The World Values Survey[23] is an international research programme devoted to the scientific and academic study of social, political, economic, religious and cultural values across the world. It produces a set of questions that are used to measure the cultural values of any country or society.

Human preferences in each AI application area:

– Many AI application systems – such as conversational agents, health assistants and online learning agents – are designed to have certain human traits. These are based on user studies and human–computer interactions (HCI) research. There is often industry consensus as to what they should be for a particular area of application. For example, health assistants that interact with patients are commonly designed to be empathetic, friendly, helpful and careful but not pedantic; and to use layperson's terms (plain language) rather than specialist medical terms. Customer service chatbots are designed to be friendly, helpful and informative and they may nudge or offer recommendations. When designing such AI application systems, it is advisable to adopt the industry-standard practice of incorporating user studies and expert advice. In each case, the choice of such attributes should be made explicit and transparent. When an AI system is trained end to end from data – such as customer service data – these attributes may be learned implicitly rather than explicitly. In such cases, the provenance of training data should be disclosed.

2.2 | Design for values

The design for values approach involves embedding human values into the AI design process through the systematic integration of ethical considerations, transparency and rigorous evaluation methods. This method has been proposed by various groups and authors who emphasize the importance of incorporating ethical considerations throughout the technological design process and directly instituting human values into the operational processes of AI systems.[24] Practical approaches to design for values extend this framework by focusing on the systematic incorporation of values into engineering and design practices. This work emphasizes the importance of considering values such as sustainability, privacy and safety from the

earliest stages of the design process, ensuring that these values are not only considered but actively shaped through design choices.[25] Methods such as the "glass box"[26] provide strategies for allowing AI decision-making processes to be transparent and interpretable, aligning them with the principles of accountability and trustworthiness. This method details how to map moral values into explicit verifiable norms that constrain the inputs and outputs of AI systems, so that adherence to these values through continuous monitoring is guaranteed. Figure 4 provides examples of how values are addressed through different design approaches (see also Figure 2).

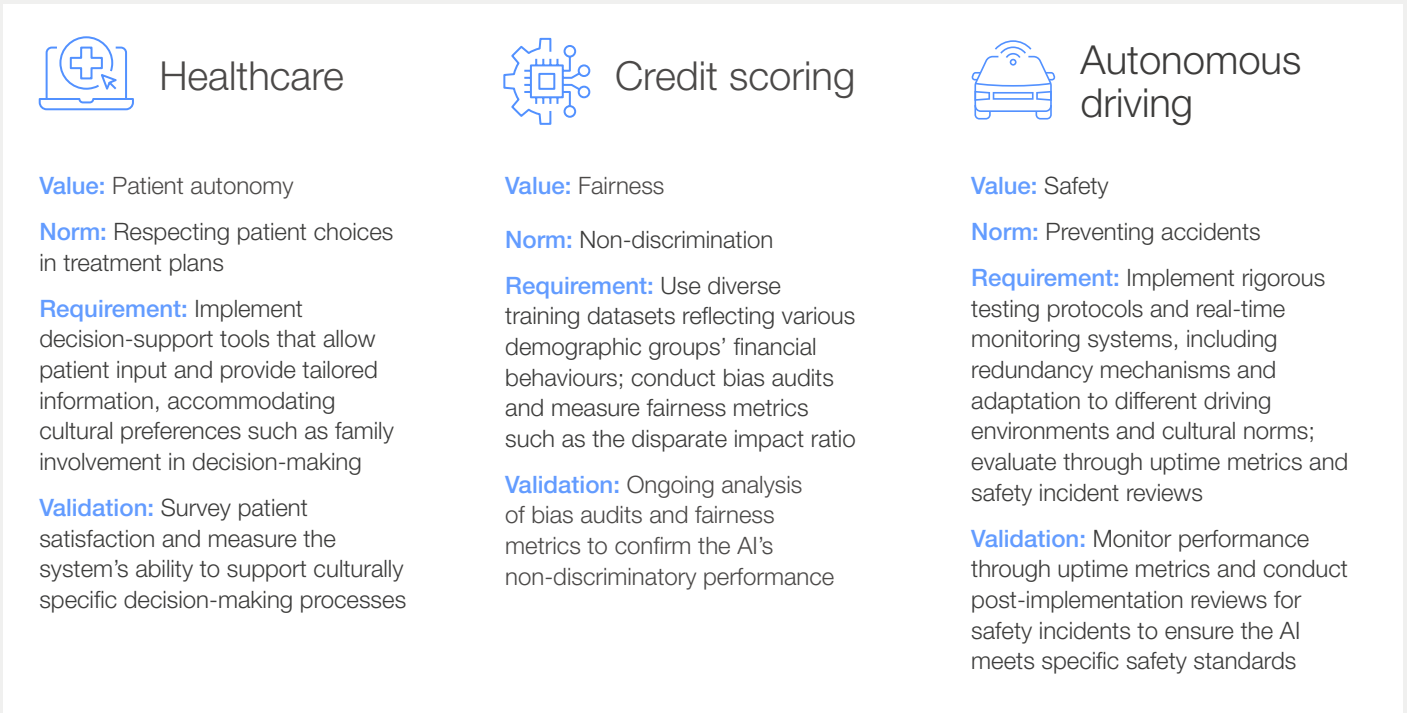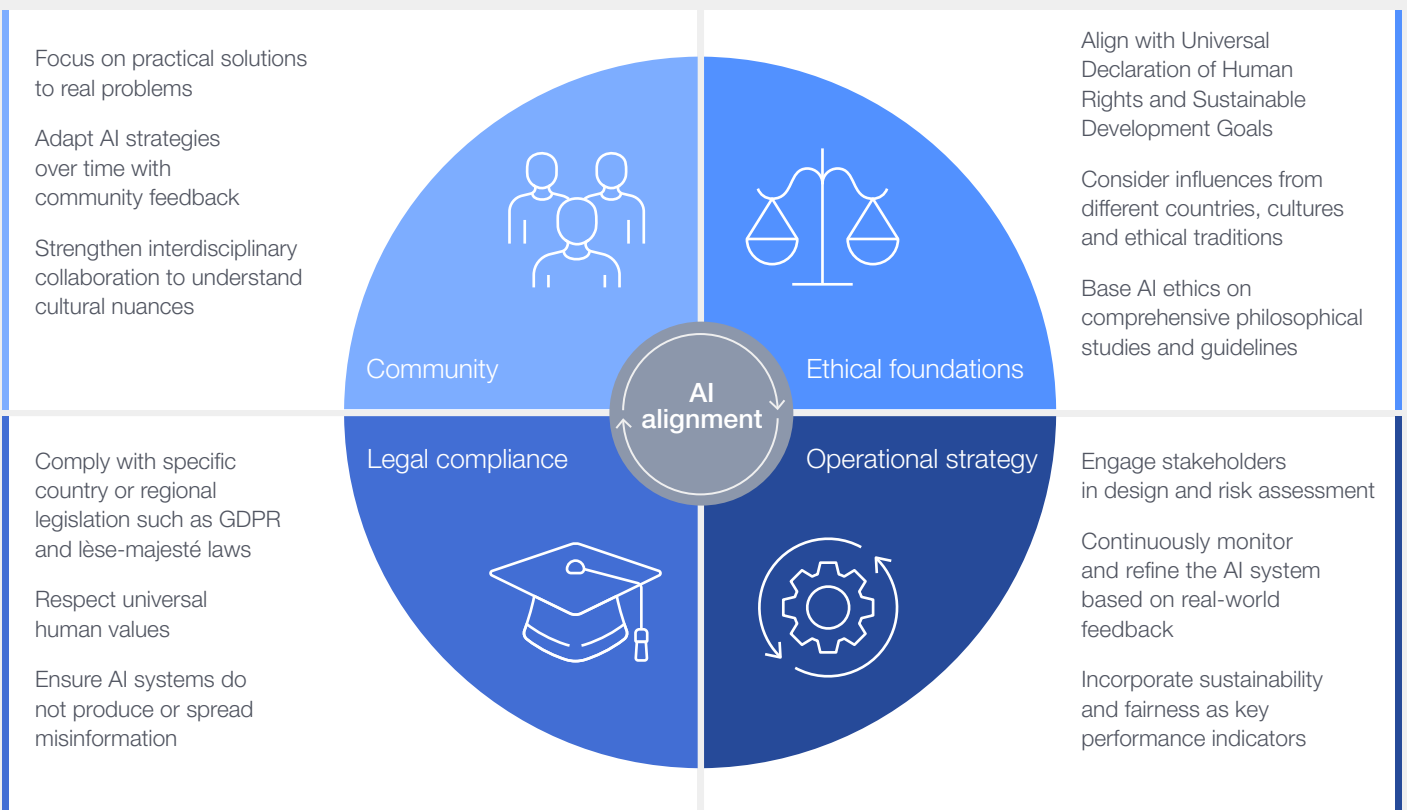FIGURE 4 | Design for values in different sectors

Healthcare

Value: Patient autonomy

Norm: Respecting patient choices in treatment plans

Requirement: Implement decision-support tools that allow patient input and provide tailored information, accommodating cultural preferences such as family involvement in decision-making

Validation: Survey patient satisfaction and measure the system's ability to support culturally specific decision-making processes

Credit scoring

Value: Fairness

Norm: Non-discrimination

Requirement: Use diverse training datasets reflecting various demographic groups' financial behaviours; conduct bias audits and measure fairness metrics such as the disparate impact ratio

Validation: Ongoing analysis of bias audits and fairness metrics to confirm the AI's non-discriminatory performance

Autonomous driving

Value: Safety

Norm: Preventing accidents

Requirement: Implement rigorous testing protocols and real-time monitoring systems, including redundancy mechanisms and adaptation to different driving environments and cultural norms; evaluate through uptime metrics and safety incident reviews

Validation: Monitor performance through uptime metrics and conduct post-implementation reviews for safety incidents to ensure the AI meets specific safety standards

Figure 5 outlines how elements of the value alignment process combine in considerations from ethical foundations to operational strategy, legal compliance and community-based concerns. The figure should be viewed as a non-exhaustive illustration of how the value alignment process needs to incorporate values at multiple levels, which should all be considered when developing and releasing new AI products and systems.

FIGURE 5 | **Select elements in value alignment**

Community
- Focus on practical solutions to real problems
- Adapt AI strategies over time with community feedback
- Strengthen interdisciplinary collaboration to understand cultural nuances

Ethical foundations
- Align with Universal Declaration of Human Rights and Sustainable Development Goals
- Consider influences from different countries, cultures and ethical traditions
- Base AI ethics on comprehensive philosophical studies and guidelines

Legal compliance
- Comply with specific country or regional legislation such as GDPR and lèse-majesté laws
- Respect universal human values
- Ensure AI systems do not produce or spread misinformation

Operational strategy
- Engage stakeholders in design and risk assessment
- Continuously monitor and refine the AI system based on real-world feedback
- Incorporate sustainability and fairness as key performance indicators

(Centre: AI alignment)

Source: World Economic Forum

2.3 | The value alignment process

Existing guidelines advocate for the continuous integration of responsibility principles throughout the AI life cycle to ensure ethical and trustworthy development. These principles promote earned trust in AI systems and enhance understanding of how AI technologies are used. This means that development and deployment processes should be explainable, transparent and robust. Organizations' use or application of AI should, where possible, be done in ways that align with these principles, thus helping to build trust and confidence in AI and enabling humans to exercise their self-determination and discretionary power. In essence, responsibility principles need to be integrated across the entire AI development life cycle, justifying decisions from ideation to usage.

TABLE 1 | Concrete actions to implement human values

Phase	Core human values	Concrete actions
Conception and analysis	**Identify key values:** Engage stakeholders to identify key human values such as fairness, transparency and privacy. **Value interpretation:** Translate these values into specific, actionable norms. For example, fairness can be interpreted as providing equal access to services for all user groups.	**Risk assessment:** Conduct a risk assessment to identify potential ethical and societal impacts of the AI system. **Stakeholder involvement**: Ensure continuous involvement of diverse stakeholders to capture a wide range of perspectives and requirements.
Design and development	**Value embedding:** Implement design strategies that incorporate identified human values into the system architecture. For instance, use bias mitigation techniques to uphold fairness. **Value validation:** Develop prototypes and conduct user testing to validate that the system aligns with the core human values identified.	**Technical specifications:** Define and implement secure data-handling practices to maintain privacy. **Performance metrics:** Establish and monitor metrics to ensure that the system performs according to the established human values.
Deployment, operation and use	**Continuous monitoring**: Implement mechanisms for ongoing monitoring of the system to ensure it continues to align with core human values. **User feedback:** Collect and incorporate user feedback to address any emerging value-alignment issues.	**System robustness:** Check the system remains robust and reliable under a variety of conditions. **Compliance audits:** Regularly audit the system to monitor compliance with relevant regulations and standards.

3 | Enablers for value alignment

Continuous stakeholder engagement is needed to ensure that AI systems are developed and deployed responsibly.

This section explores important enablers of value alignment, which include frameworks and guidelines, human engagement, organizational change, and audits and assessments. Each plays a critical role in embedding ethical principles into the life cycle of AI systems, from design and development to deployment and operation. By making use of these and other enablers, stakeholders can build and release AI systems that are effective and innovative as well as ethical and trustworthy.

3.1 | Frameworks and guidelines

Frameworks and guidelines are essential in value alignment as they provide structured approaches and best practices to make sure AI systems operate in ways that are ethically sound and align with human values. These tools serve as foundational elements to guide AI technologies' development, deployment and management, evaluating whether they meet established ethical standards and societal expectations. By implementing these frameworks, organizations can navigate the complex landscape of AI ethics, enhance transparency and promote trust among users and stakeholders.

3.2 | Human engagement

Human engagement in AI is a continuous and dynamic process that plays a vital role in the iterative refinement and enhancement of AI systems. This ongoing process builds on initial multistakeholder consultations, emphasizing continuous feedback and adaptation to maintain relevance and trust. Key to this process is employing techniques such as inverse reinforcement learning,[27] whereby AI systems learn from observing human behaviour, and adopting value-based approaches that prioritize intrinsic human values. Moreover, participatory design empowers users, including members of marginalized communities, to influence AI development actively, ensuring that the technology addresses diverse needs and ethical considerations.

Human engagement throughout the whole life cycle keeps AI systems aligned with evolving ethical standards and societal expectations by integrating real-world experiences and adapting the systems to feedback. Techniques such as inverse reinforcement learning and value-based approaches evolve with the systems, so alignment with changing human goals and values can be maintained, thus ensuring relevance and trustworthiness over time.

User-centred design is critical, particularly in sensitive sectors such as healthcare, in which input from various backgrounds enhances an AI system's fairness and inclusiveness and minimizes its biases. AI literacy initiatives are vital for all stakeholders, facilitating informed participation in AI development and policy-making. Ethical review boards and continuous feedback mechanisms are essential for upholding ethical standards and enabling AI systems to evolve constructively in response to societal feedback.

3.3 | Organizational change

Organizational change is a pivotal strategy for ensuring value alignment in AI, focusing on reshaping organizational culture, processes and policies to promote ethical AI integration. This approach views the adaptation of organizational structures as crucial for embedding value alignment principles, so that AI systems continue to reflect and uphold both internal values and societal ethical standards, and includes the following key elements:

– **Cultural adaptation:** Cultivate a culture that prioritizes ethical considerations in AI deployment, encouraging a mindset shift towards responsible AI use.

– **Governance frameworks:** Implement robust governance structures to enforce transparency, accountability and ethical compliance in AI initiatives and align them with organizational and societal values.

– **Education and training:** Reinforce that continuous learning programmes are vital to enhance AI ethics awareness among employees, equipping them to identify and mitigate biases in AI systems and practices.

– **Policy integration:** Ensure organizational policies explicitly reflect value alignment principles, guiding the development, deployment and management of AI technologies so that they are ethically aligned and socially responsible.
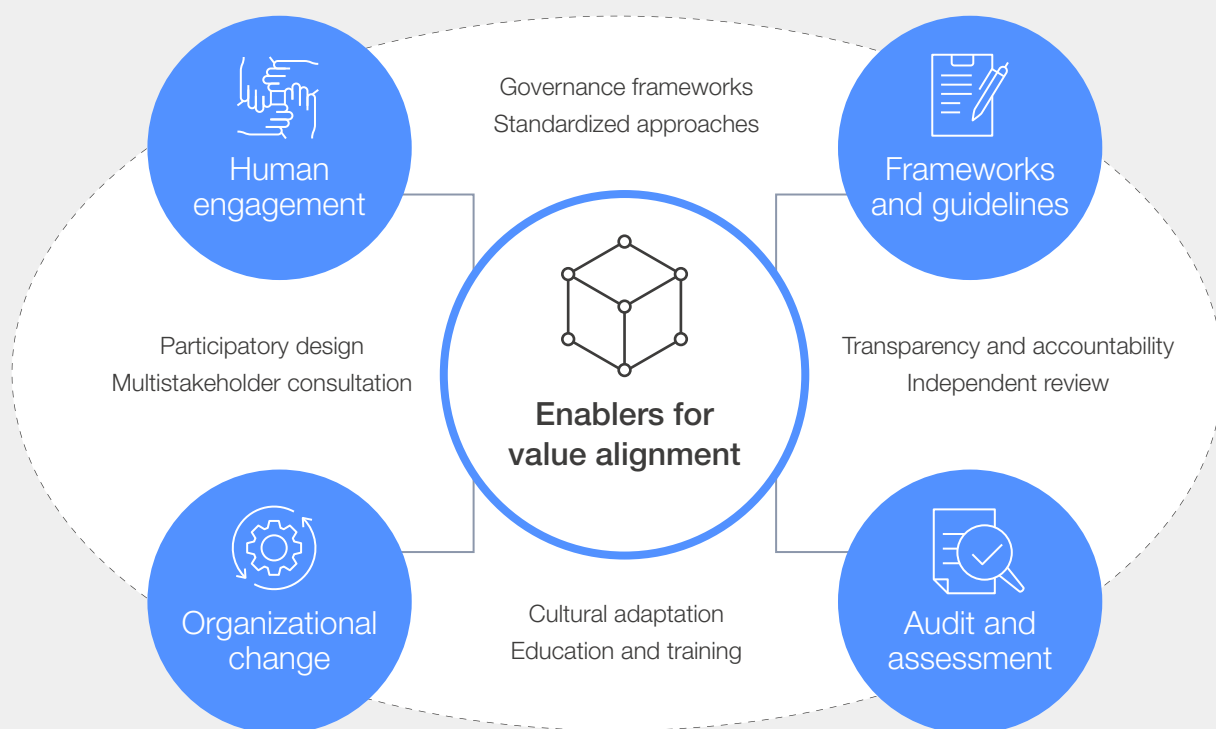
One way to achieve the organizational change required is to adhere to standards such as ISO/IEC 42001,[28] which outlines the criteria for setting up, executing, maintaining and enhancing AI management systems (AIMs) in companies. By following existing standards, organizations that offer or use AI-driven products or services should be better equipped to guarantee the responsible development and application of AI systems. Aligning further with the guidance provided by ISO/IEC JTC 1/SC 42,[29] which focuses on standardization in AI, can further support this. A holistic approach such as this involves reshaping organizational culture through training and awareness programmes, redesigning processes and policies to enforce ethical standards and implementing continuous monitoring and improvement mechanisms to ensure AI systems remain aligned with ethical standards over time.

3.4 | Audits and assessments

Audits and assessments are critical in making sure AI systems consistently adhere to established value alignment principles and ethical standards. These processes provide a structured framework for evaluating the effectiveness of AI systems in maintaining alignment with human values and societal norms over time. Through systematic audits and assessments, organizations can proactively manage their AI systems' ethical integrity and societal impact, ensuring they remain aligned with value alignment objectives throughout their life cycle. Core aspects include:

– **Independent review:** Establish independent audits to ensure objectivity, allowing for unbiased evaluation of AI systems against ethical and value alignment benchmarks.

– **Non-independent review**: Arrange these to be conducted by internal teams, which offer context-specific insights, continuous feedback loops, early issue detection and stakeholder engagement, and which complement independent audits that check comprehensive ethical alignment and action timely mitigation of potential issues in AI systems.

– **Technical expertise:** Engage auditors with specialized knowledge in AI and ethics, who are essential to accurately assess complex AI systems and identify areas for improvement.

– **Shared standards:** Adopt shared national and international standards to facilitate a consistent and transparent approach to evaluating AI systems, allowing for comparability and benchmarking across different entities and sectors.

– **Explainability:** Assess the explainability of AI systems, which is crucial to ensure stakeholders can understand decision-making processes and outcomes, promoting trust and facilitating value alignment.

– **Comprehensive evaluation:** Conduct audits that cover all aspects of AI systems, including design, development, deployment and ongoing operation so that comprehensive value alignment can be maintained.

– **Continuous monitoring:** Check that assessment findings lead to actionable insights, driving continuous improvement in AI systems to align with evolving ethical and societal expectations.

– **Transparency and accountability:** Ensure audit and assessment results are transparently reported to promote trust and accountability among stakeholders, including users, regulators and the public.

FIGURE 6 | Enablers for value alignment

Governance frameworks
Standardized approaches

Human engagement

Frameworks and guidelines

Participatory design
Multistakeholder consultation

Enablers for value alignment

Transparency and accountability
Independent review

Organizational change

Cultural adaptation
Education and training

Audit and assessment

(4) Values and red lines

AI value alignment and red lines are intrinsically related, helping to ensure that AI systems adhere to ethical standards and harmful outcomes are prevented.

The value alignment process is intrinsically connected to the concept of red lines in AI ethics, which represent non-negotiable boundaries that AI systems must not cross. In defining the limits of acceptable AI behaviour, red lines ensure that AI systems adhere to fundamental ethical standards and do not compromise human rights and societal norms. For instance, a red line might prohibit AI systems from making autonomous decisions that could result, without human oversight, in harm to individuals or allow the systems to engage in discriminatory practices. To uphold these red lines effectively, it is crucial to embed core human values such as fairness, transparency and privacy into the design and operational phases to prevent unethical behaviour and maintain trust and safety in AI deployment.

The establishment of global red lines is essential to provide a universal ethical framework that all AI systems must follow, regardless of where they are developed or deployed. These ensure that fundamental human rights and ethical standards are upheld worldwide, creating a baseline of trust and safety. However, while such universal standards are critical, local contexts and specific applications of AI may require tailored red lines that address the unique ethical challenges and societal expectations linked to their context of application. Therefore, steps are necessary to establish specific, local red lines tailored to each situation, including:

1. **Systematic risk assessments:** Conduct thorough risk assessments to identify ethical impacts, define red lines clearly and develop strategies to prevent crossing these boundaries.

2. **Stakeholder involvement:** Engage diverse stakeholders to reach consensus on red lines so that AI aligns with societal values.

3. **Compliance audits:** Regularly conduct independent audits to verify adherence to ethical standards and red lines, maintaining detailed records of compliance checks.

4. **Continuous monitoring:** Implement real-time monitoring and alert systems to track AI behaviour, allowing immediate human intervention if a red line is breached.

5. **Robust validation processes:** Establish pre-deployment testing to ensure AI systems do not cross red lines and develop metrics to measure adherence.

6. **Transparency and explainability:** Provide detailed documentation on how red lines are integrated and make sure stakeholders understand their enforcement.

7. **Human oversight and accountability:** Maintain human responsibility for checking that AI systems do not cross red lines, and offer clear intervention protocols.

8. **Adaptive governance:** Develop frameworks that evolve with technological advances and changing societal values, regularly updating red lines and involving stakeholders in the process.

By clearly defining and strictly enforcing red lines, AI systems can prevent unethical behaviour, maintaining trust and safety in their deployment. This approach ensures AI systems are ethically sound and operate within moral and legal frameworks. Ultimately, the responsibility for enforcing red lines lies with the people who design, develop and deploy AI systems, emphasizing the critical role of human oversight and accountability in responsible AI development. Continuous improvement and stakeholder engagement are essential for sustaining AI alignment with human values, fostering a future in which AI technologies enhance societal well-being.

Conclusion

By balancing operational attributes with fundamental human values, developers can create AI systems that are both reliable and ethically sound.

AI value alignment is a multifaceted and dynamic process that requires the integration of ethical principles into every stage of the AI life cycle.

Important enablers – such as frameworks, human engagement, organizational change and audits – provide structured approaches to embed values into AI systems. These tools and methodologies ensure that AI systems are developed, deployed and managed in ways that uphold human dignity, protect individual rights and foster mutual trust and transparency.

By continuously involving stakeholders, conducting rigorous audits and maintaining flexibility to accommodate diverse cultural interpretations of values, organizations can work to keep AI systems aligned with societal norms and ethical standards. The paper underscores the critical connection between value alignment and AI ethics red lines, emphasizing that respecting these boundaries is essential to prevent unethical behaviour and maintain trust in AI technologies.

The pursuit of value alignment in AI may be a technical challenge but it is a societal imperative, requiring ongoing collaboration, transparency and accountability among stakeholders. This comprehensive approach will help AI technologies contribute positively to society while adhering to the highest ethical standards.

Appendix: Resources

The section contains principles, frameworks, guidelines and other helpful resources to assist in the AI value alignment process.

National and international governments and agencies

– AI Verify Foundation. (n.d.). *Model AI governance framework for generative AI*. Retrieved August 21, 2024, from https://aiverifyfoundation.sg/resources/mgf-gen-ai/

– China Electronics Standardization Institute. (2021, October 21). *Original CSET translation of artificial intelligence standardization white paper.* https://cset.georgetown.edu/publication/artificial-intelligence-standardization-white-paper-2021-edition/

– European Union. (2024). EU AI Act 2024: *First regulation on artificial intelligence*. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

– European Union. (2022, June 7). *High-level expert group on artificial intelligence*. https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai#:~:text=AI%20HLEG%20and%20the%20European,upcoming%20legislative%20steps%20in%20AI

– National Institution for Transforming India. (2018, June). *National strategy for artificial intelligence*. https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf

– Organisation for Economic Co-operation and Development (OECD). (2019). *Recommendation of the Council on Artificial Intelligence*. https://www.fsmb.org/siteassets/artificial-intelligence/pdfs/oecd-recommendation-on-ai-en.pdf

– Singapore Digital. (2020). *Model artificial intelligence governance framework*. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

– United Kingdom government. (2020). *Data ethics framework*. https://www.gov.uk/government/publications/data-ethics-framework

– United Nations Children's Fund (UNICEF). (2021, November). *Policy guidance on AI for children*. https://www.unicef.org/globalinsight/media/2356/file/UNICEF-Global-Insight-policy-guidance-AI-children-2.0-2021.pdf

– United Nations Educational, Scientific and Cultural Organization (UNESCO). (2021, November 23). *Recommendation on the ethics of artificial intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000381137?3=null&queryId=c5dd8ced-9647-452b-b4d6-92723006496c

– United States Federal Data Strategy. *Federal data strategy: Data ethics framework*. https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf

– White House. (2023, October 30). E*xecutive order on the safe, secure, and trustworthy development and use of artificial intelligenc*e. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

– World Health Organization. (2024, January 18). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. https://www.who.int/publications/i/item/9789240084759

Private companies, non-governmental and non-profit organizations

– Fairlearn. (n.d.). *Improve fairness of AI systems*. Retrieved August 21, 2024, from https://fairlearn.org/

– Fenech, M., Strukelj, N., & Buston, O. (2018). *Ethical, social, and political challenges of artificial intelligence in health*. Wellcome Trust/Future Advocacy. https://wellcome.org/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf

– Floridi, L., et al. (2018, November 26). *AI4People – an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds & Machines*, 28, 689–707. https://link.springer.com/article/10.1007/s11023-018-9482-5#citeas Future of Life Institute. (2017, August 11). Asilomar AI principles. https://futureoflife.org/open-letter/ai-principles/

– IBM. (n.d.). *AI Fairness 360: An extensible open-source toolkit*. Retrieved August 21, 2024, from https://github.com/Trusted-AI/AIF360

– Internet Society. (2017, April 18). *Artificial intelligence and machine learning: Policy pape*r. https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/

– Microsoft. (n.d.). *Empowering responsible AI practices*. Retrieved August 21, 2024, from https://www.microsoft.com/en-us/ai/responsible-ai / https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl?culture=en-us&country=us

– Partnership on AI. (n.d.). *PAI brings together a diverse community to address important questions about our future with AI*. Retrieved August 21, 2024, from https://www.partnershiponai.org

– System Analysis Program Development (SAP). (n.d.). *SAP's guiding principles for artificial intelligence*. Retrieved August 21, 2024, from https://www.sap.com/documents/2018/09/940c6047-1c7d-0010-87a3-c30de2ffd8ff.html

– Telefónica. (n.d.). *Telefónica's approach to the responsible use of AI*. Retrieved August 21, 2024, from https://www.telefonica.com/en/wp-content/uploads/sites/5/2021/08/ia-responsible-governance.pdf

Academic institutions

– Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

– Université de Montréal. (2017, November 3). The University of Montreal declaration for responsible development of artificial intelligence. https://recherche.umontreal.ca/english/strategic-initiatives/montreal-declaration-for-a-responsible-ai/

– Whittaker, M., et al. (2018, December). *AI Now report 2018*, 1–62. AI Now Institute, New York University. https://ec.europa.eu/futurium/en/system/files/ged/ai_now_2018_report.pdf

Professional/standards associations

– Shahriari, K., & Shahriari, M. (2017). *IEEE standard review – ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems*. 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), 197–201. https://www.semanticscholar.org/paper/IEEE-standard-review-%E2%80%94-Ethically-aligned-design%3A-A-Shahriari-Shahriari/9fea647551c280d517f5a478c4212868b8cddc9

Contributors

Global Future Council on the Future of AI 2023–2024

The World Economic Forum's Network of Global Future Councils is the world's foremost multistakeholder and interdisciplinary knowledge network dedicated to promoting innovative thinking to shape a more resilient, inclusive and sustainable future.

Global Future Council on the Future of AI members

Co-Chair

Virginia Dignum
Professor of Responsible Artificial Intelligence, Umeå University

Members

Fatmah Baothman
Chief Executive Officer, Abdullah Alothain AI and R&D

Karim Beguir
Co-Founder and Chief Executive Officer, InstaDeep

Saqr Binghalib
Executive Director, Artificial Intelligence, Digital Economy and Remote Work Applications Office at the Prime Minister's Office, United Arab Emirates

Vilas Dhar
President and Trustee, Patrick J. McGovern Foundation

Jibu Elias
Country Lead, India, Responsible Computing Challenge, Mozilla Foundation

Kay Firth-Butterfield
Senior Research Fellow, University of Texas at Austin

Alice Friend
Global Head, Artificial Intelligence and Emerging Tech Policy, Google

Pascale Fung
Chair; Professor, Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology

Constanza Gomez-Mont
Founder and Chief Executive Officer, C Minds

Sara Hooker
Vice-President Research, Head of Cohere for AI

Mohan Kankanhalli
Provost's Chair Professor of Computer Science, National University of Singapore

Edson Prestes e Silva Júnior
Full Professor, Informatics Institute, Federal University of Rio Grande do Sul

Francesca Rossi
IBM Fellow; Global Leader, Artificial Intelligence Ethics, IBM

Adrian Weller
Director of Research, Machine Learning, University of Cambridge

World Economic Forum

Council Managers

Talal Altook
Fellow, Artificial Intelligence and Machine Learning

Benjamin Cedric Larsen
Lead, Artificial Intelligence and Machine Learning

Hannah Rosenfeld
Specialist, Artificial Intelligence and Machine Learning

Acknowledgements

Production

Endnotes

1. Dignum, V., & Dignum, F. (Eds.). (2013). *Perspectives on culture and agent-based simulations: Integrating cultures.* Springer. https://doi.org/10.1007/978-3-319-01952-9

2. ScienceDirect. (2023). *Disparate impact.* https://www.sciencedirect.com/topics/computer-science/disparate-impact#:~:text=Disparate%20Impact%20refers%20to%20the,higher%20bias%20in%20the%20predictions

3. Uptime is a metric that measures system reliability and stability. It is represented as the percentage of time that a system remains operational and is available to use.

4. Arora, S., & Doshi, P. (2021, August). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, *297*, 103500. https://doi.org/10.1016/j.artint.2021.103500

5. Lee, H., et al. (2023). Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*. https://arxiv.org/abs/2309.00267

6. Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems, 35,* 27730–27744. https://arxiv.org/abs/2203.02155

7. United Nations. (1948). *Universal Declaration of Human Rights*. https://www.un.org/en/about-us/universal-declaration-of-human-rights

8. United Nations. *The 17 Goals*. https://sdgs.un.org/goals

9. American Red Cross. (2011). *Summary of the Geneva Conventions of 1949 and their Additional Protocols*. https://www.redcross.org/content/dam/redcross/atg/PDF_s/International_Services/International_Humanitarian_Law/IHL_SummaryGenevaConv.pdf

10. Council of Europe. (2024, May 17). *Council of Europe adopts first international treaty on artificial intelligence*. www.coe.int/en/web/portal/-/council-of-europe-adopts-first-international-treaty-on-artificial-intelligence

11. Organisation for Economic Co-operation and Development (OECD). (n.d.). *OECD AI Principles overview*. Retrieved August 21, 2024, from https://oecd.ai/en/ai-principles

12. Singer, P. (2024, August 5). The history of western ethics. *Encyclopedia Britannica*. https://www.britannica.com/topic/ethics-philosophy/The-history-of-Western-ethics

13. Landolt, H. (1999). Henry Corbin, 1903–1978: Between philosophy and orientalism. *Journal of the American Oriental Society*, *119*(3), 484–490. https://doi.org/10.2307/605942

14. Gyekye, K. (2010, September 9). African ethics. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/african-ethics/

15. European Union. (2021). *Regulation (EU) 2021/0106 of the European Parliament and of the Council*. Official Journal of the European Union. https://artificialintelligenceact.eu/wp-content/uploads/2024/01/AI-Act-FullText.pdf

16. Ministry of Foreign Affairs of the People's Republic of China. (2022, November 17). *Position paper of the People's Republic of China on strengthening ethical governance of artificial intelligence (AI)*. https://www.fmprc.gov.cn/eng/zy/wjzc/202405/t20240531_11367525.html

17. Zhang, L. (2019, September 9). *China: AI Governance Principles released*. Library of Congress. https://www.loc.gov/item/global-legal-monitor/2019-09-09/china-ai-governance-principles-released/

18. US Office of Science and Technology Policy. (2022, October). *Blueprint for an AI bill of rights: Making automated systems work for the American people*. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

19. Singapore Statutes Online. (2019). *Protection from Online Falsehoods and Manipulation Act 2019 (Act No. 18 of 2019)*. https://sso.agc.gov.sg/Act/POFMA2019

20. European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament, Article 17*. https://gdpr-info.eu/

21. DLA Piper (2023, January 29). *Data protection laws of the world: United States*. https://www.dlapiperdataprotection.com/index.html?c=US&c2=&go-button=GO&t=law

22. European Union. (2021). *Regulation (EU) 2021/0106 of the European Parliament and of the Council*. https://artificialintelligenceact.eu/wp-content/uploads/2024/01/AI-Act-FullText.pdf

23. World Values Survey Association. *World Values Survey*. https://www.worldvaluessurvey.org/wvs.jsp

24. Friedman, B., & Kahn, P. H. (2003). *Value sensitive design: Shaping technology with moral imagination*. MIT Press. https://direct.mit.edu/books/monograph/4328/Value-Sensitive-DesignShaping-Technology-with; Flanagan, M., & Nissenbaum, H. (2014). *Values at play in digital games.* MIT Press. https://direct.mit.edu/books/monograph/4030/Values-at-Play-in-Digital-Games

25. TU Delft Design for Values Institute. (n.d.). *Design for values: Research, education & collaboration*. Retrieved August 21, 2024, from https://www.delftdesignforvalues.nl/#:~:text=The%20Delft%20Design%20for%20Values,Electric%20Engineering%2C%20Mathematics%2C%20and%20Computer

26. Dignum, V., et al. (2019). Governance by glass-box: Implementing transparent moral bounds for AI behaviour. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 5787–579*3. https://www.ijcai.org/Proceedings/2019/802

27. Arora, S., & Doshi, P. (2021, August). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence, 297*, 103500. https://doi.org/10.1016/j.artint.2021.103500

28. International Organization for Standardization. (2023). *ISO/IEC 42001:2023: Information technology – artificial intelligence – management system*. https://www.iso.org/standard/81230.html

29. International Organization for Standardization. (2017). *ISO/IEC JTC 1/SC 42: Artificial intelligence*. https://www.iso.org/committee/6794475.html