INNOVATION

AI value alignment: How we can align artificial intelligence with human values

Oct 17, 2024

AI value alignment is about ensuring that artificial intelligence (AI) systems act in accordance with shared human values and ethical principles.
Image: Getty Images/iStockphoto

Benjamin Larsen
Initiatives Lead, AI Systems and Safety, Centre for AI Excellence, World Economic Forum

inia Dignum
ssor of Responsible Artificial Intelligence, Umeå University

- **Human values are not uniform across regions and cultures, so AI systems must be tailored to specific cultural, legal and societal contexts.**
- **Continuous stakeholder engagement – including governments, businesses, and civil society – is key to shaping AI systems that align with human values.**

As AI continues to integrate into almost every aspect of life – from healthcare to autonomous driving – there is a growing imperative to ensure that AI systems reflect and uphold shared human values.

The October 2024 Global Future Council white paper, AI Value Alignment: Guiding Artificial Intelligence Towards Shared Human Goals, tackles this pressing issue, exploring how we can guide AI systems to align with societal values such as fairness, privacy and justice. This alignment is not just a technical challenge but a societal responsibility.

Have you read?

- **Why the human touch is needed to harness AI tools for communications**
- **Could fear of AI pose the biggest risk of all to humanity?**

What is AI value alignment?

AI value alignment refers to the designing of AI systems that behave in ways consistent with human values and ethical principles. However, this is easier said than done.

but its interpretation can differ greatly between regions. While some countries prioritize individual privacy, others may emphasize collective security over personal data protection.

At its core, value alignment aims to embed core human values into AI systems at every stage of development, from design to deployment. This process requires translating abstract ethical principles into practical technical guidelines and ensuring that AI systems remain auditable and transparent.

For example, an AI system used in healthcare needs to balance patient autonomy, fairness in decision-making and privacy while also being robust and compliant with regulations such as the US Health Insurance Portability and Accountability Act.

The challenge lies in operationalizing these values to make them explicit, traceable and verifiable. As highlighted in the white paper, value alignment involves continuously monitoring and updating AI systems to ensure they adapt to evolving societal norms and ethical standards.

An example of the value alignment process: AI in healthcare

To understand the practical implications of value alignment, consider an AI system used in a hospital to diagnose patients. This system must navigate key human values, such as patient autonomy and privacy. It also needs to be transparent, explaining how it arrives at its recommendations so that patients and doctors can trust its proposed suggestions and conclusions.

However, privacy and transparency can sometimes be in tension. While providing patients with detailed information fosters trust, it can also raise privacy concerns. To address this, healthcare AI systems could incorporate transparent algorithms while using encryption to protect sensitive information.

sensitive to the needs of the people it serves.

Overcoming cultural differences in value alignment

One of the key takeaways from the white paper is the importance of understanding cultural differences when developing AI systems. For example, in credit scoring, fairness might mean different things depending on the cultural context. In some societies, creditworthiness is linked to community trust and social standing; in others, it is purely a function of individual financial behaviour.

The paper advocates for a tailored approach to AI value alignment in response to these complexities. Rather than adopting a one-size-fits-all model, AI developers must consider the unique cultural, legal and societal contexts in which their AI systems operate.

For instance, a credit-scoring AI used in diverse regions might require localized training datasets that reflect the financial behaviours of different demographic groups. Auditing tools and fairness metrics, such as disparate impact ratios, can help ensure that a system does not unintentionally discriminate against any group.

Paul Daugherty

Dec 17 • The World Economic Forum Book Club Podcast

Save on Spotify

28:14

Ensuring AI value alignment requires technical innovations and organizational shifts. On the technical side, tools such as “reinforcement learning from human feedback” allow developers to integrate human values directly into AI systems. Meanwhile, value-sensitive design methods help engineers embed ethical considerations into the core architecture of AI systems from the outset.

Organizationally, achieving value alignment means fostering a culture prioritizing ethical AI development. Multi-stakeholder consultations, continuous training and the implementation of governance frameworks are essential. For example, organizations can follow standards such as ISO/IEC 42001, which outlines the criteria for setting up AI management systems, to ensure their AI products align with human values.

The role of audits and assessments

Audits are crucial in maintaining value alignment throughout the AI system’s lifecycle. Regular, independent and internal assessments ensure that AI systems continuously align with ethical standards and societal norms.

These audits should evaluate technical performance and the broader impact of AI on human rights and social equity. For instance, transparency audits help ensure that users can understand and trust the decisions made by AI systems, while fairness audits detect and mitigate bias.

> **AI can be a powerful tool for advancing societal well-being but only if we remain vigilant and align it with our shared values and principles.**
>
> njamin Larsen, Artificial Intelligence and Machine Learning Lead, World Economic Forum

The process of AI value alignment is intrinsically linked to the discussion around red lines in AI, which are the ethical boundaries that AI systems must not cross under any circumstances. These red lines provide clear moral boundaries, ensuring AI systems do not engage in harmful or unethical behaviour.

For example, a red line could prohibit AI systems from impersonating humans, engaging in unauthorized replication and breaking into other AI systems.

By establishing red lines, we can prevent AI from being used in ways that undermine human dignity or exacerbate inequality. These non-negotiable boundaries help foster trust in AI technologies, assuring that even as AI systems become increasingly powerful, they will remain ethically aligned.

A call to collective action

As AI systems become more pervasive, ensuring their alignment with human values becomes not just a technical task but a societal imperative. The Global Future Council's white paper on Value Alignment in AI provides a roadmap for achieving alignment through ethical frameworks, continuous human engagement and rigorous auditing processes.

Ultimately, the responsibility for value alignment rests not just with AI developers but with all stakeholders, from governments to businesses to civil society organizations and individuals. By fostering collaboration and transparency, we can ensure that AI systems contribute to a future where technology serves humanity's best interests and is guided by shared values.

AI can be a powerful tool for advancing societal well-being but only if we remain vigilant and align it with our shared values and principles.

Sign up for free

Stay up to date:

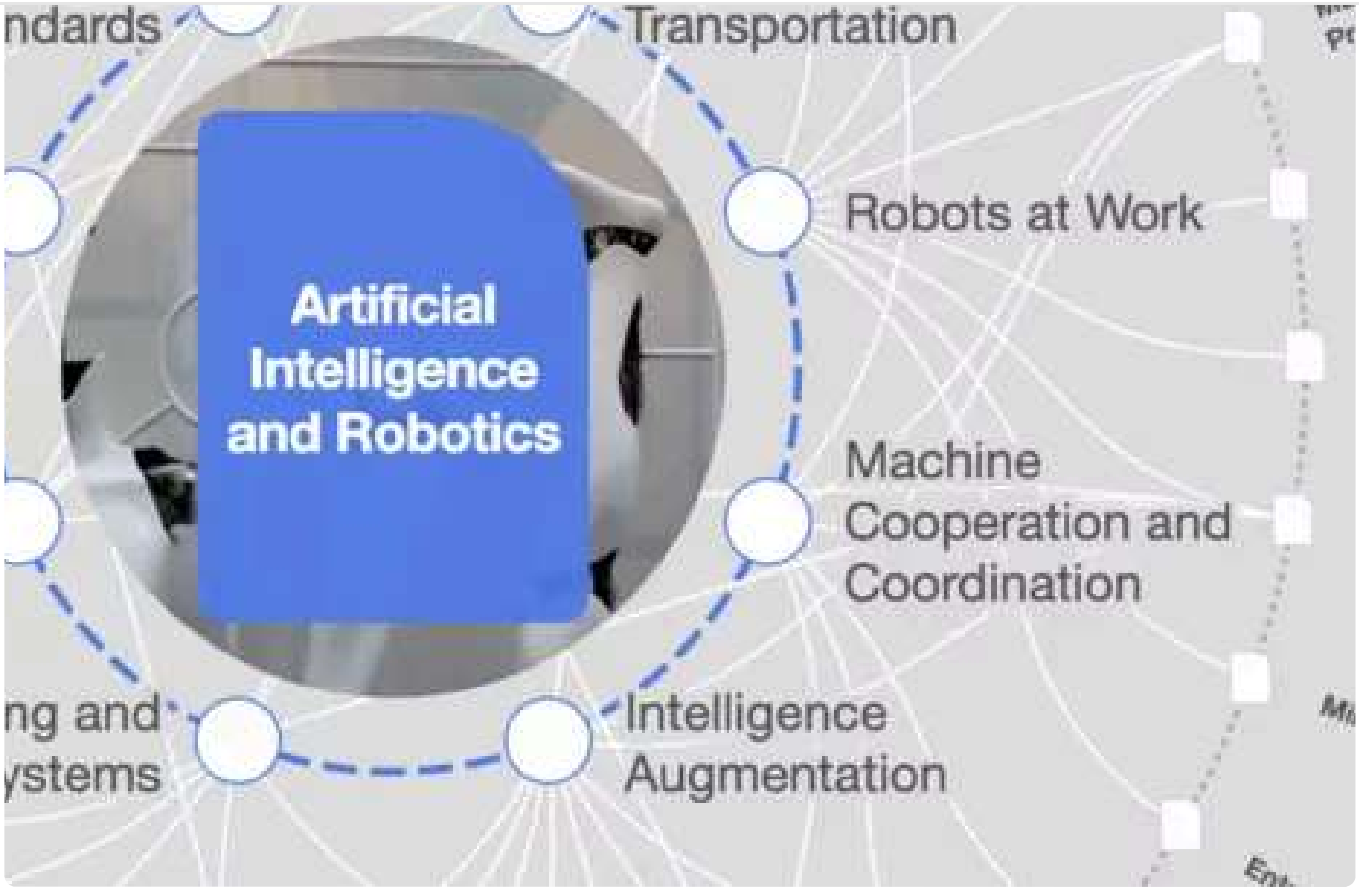
Artificial Intelligence

Follow +

Related topics:

Innovation | Forum in Focus

Share:



THE BIG PICTURE

Explore and monitor how Artificial Intelligence is affecting economies, industries and global issues

Forum Stories newsletter

Bringing you weekly curated insights and analysis on the global issues that matter.

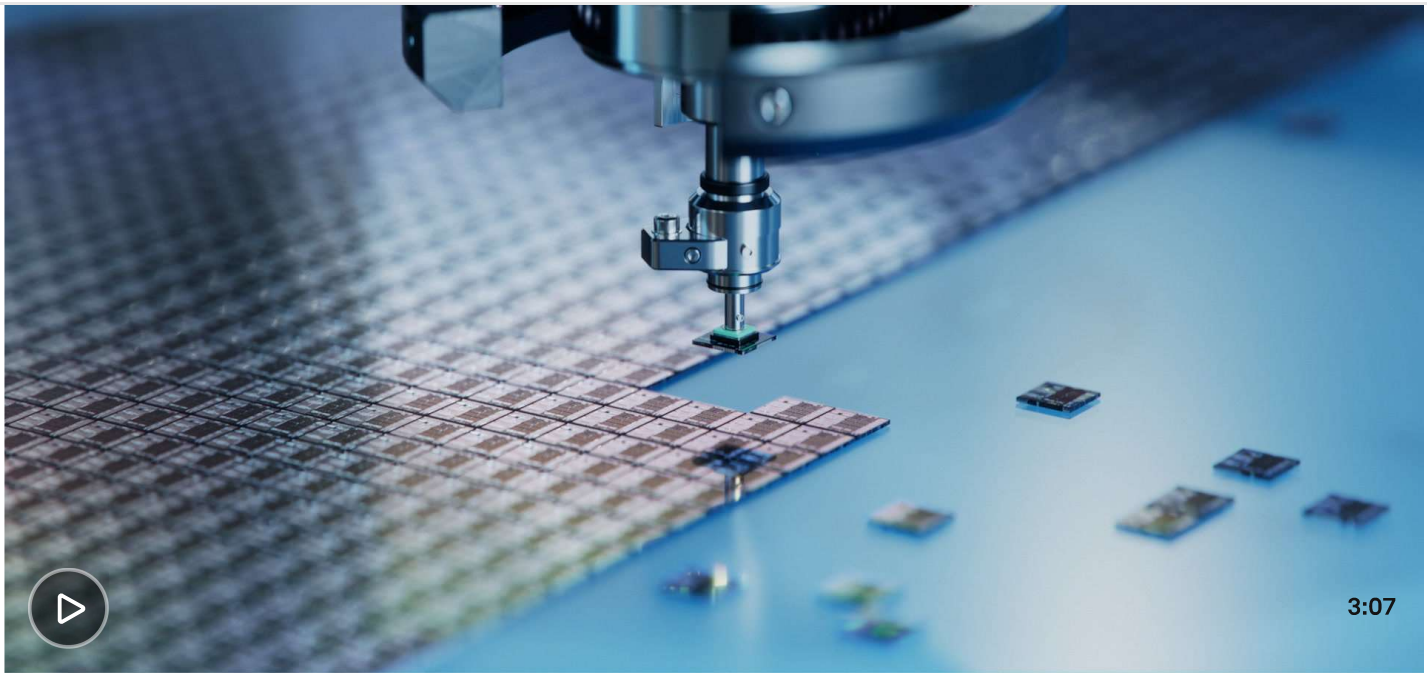
More on **Innovation**

SEE ALL

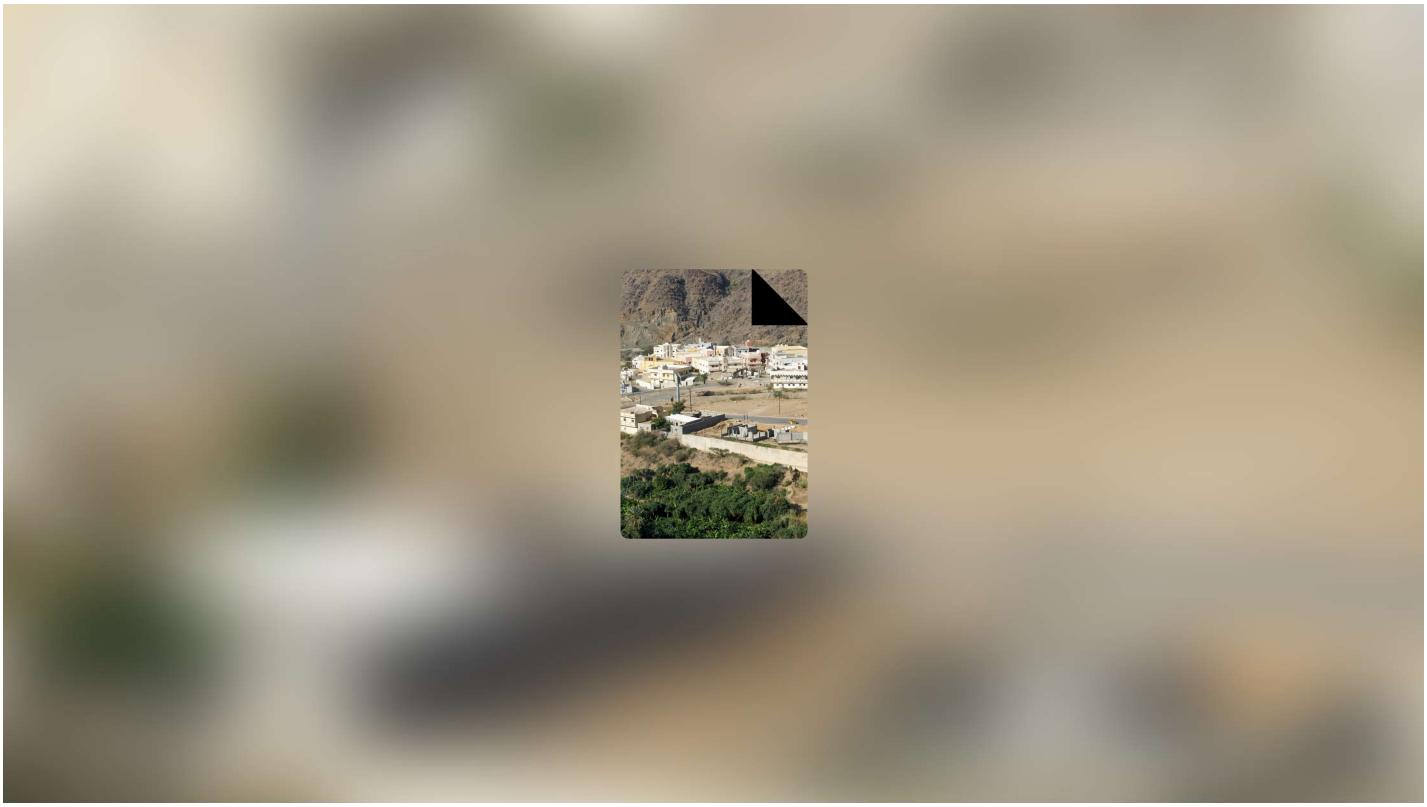
What will it take for India to lead on deep tech and the global space race?

Awais Ahmed and Srishti Bajpai

November 11, 2025

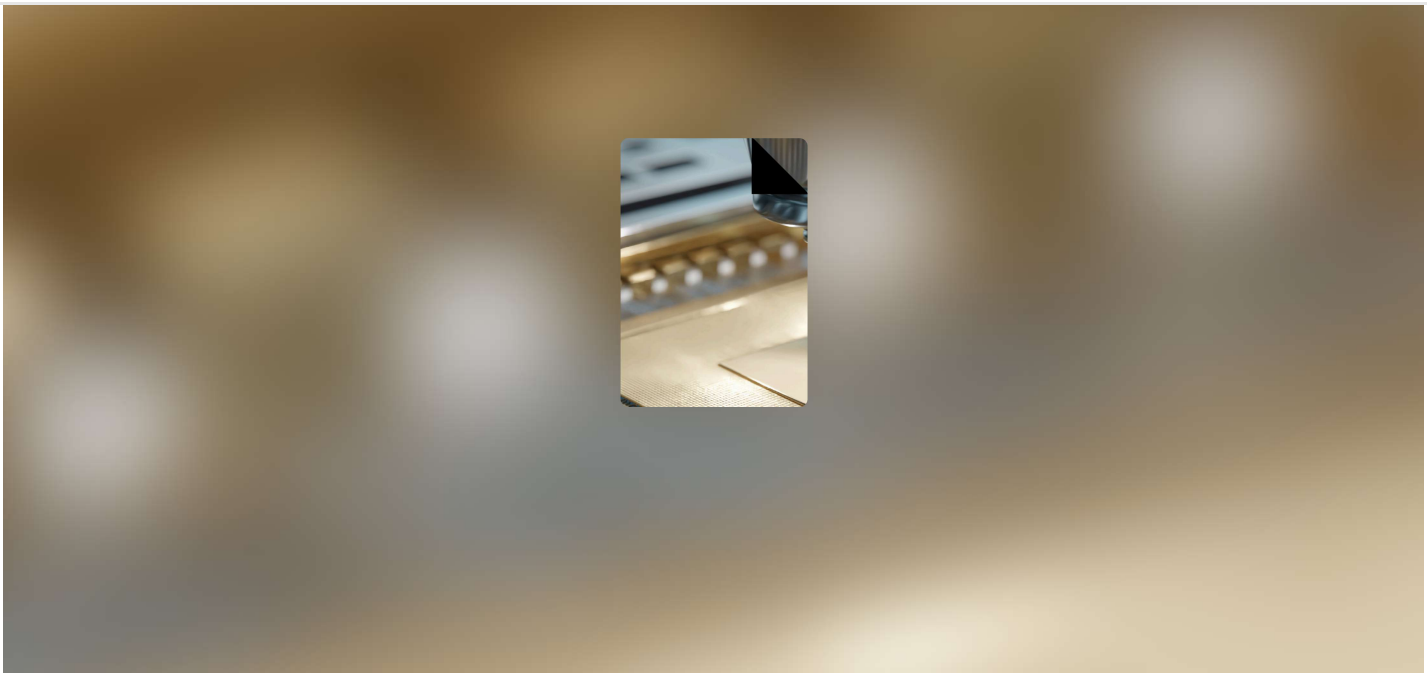


25 Years of Technology Pioneers



Deployment Pathways for Advanced Air Mobility: Lessons from Early Implementation in Saudi Arabia

29, 2025



Quantum Technologies: Key Opportunities for Advanced Manufacturing and Supply Chains

Oct 28, 2025

ing measurement into momentum so agile governance can keep pace with AI

Kelly Ommundsen

What is electrotech and what will it mean for geopolitics and energy security?

Sam Butler-Sloss and Daan Walter

October 22, 2025

About us

Our mission

Our Institutional Framework

History

Leadership and governance

Our Impact

More from the Forum

tres

Meetings