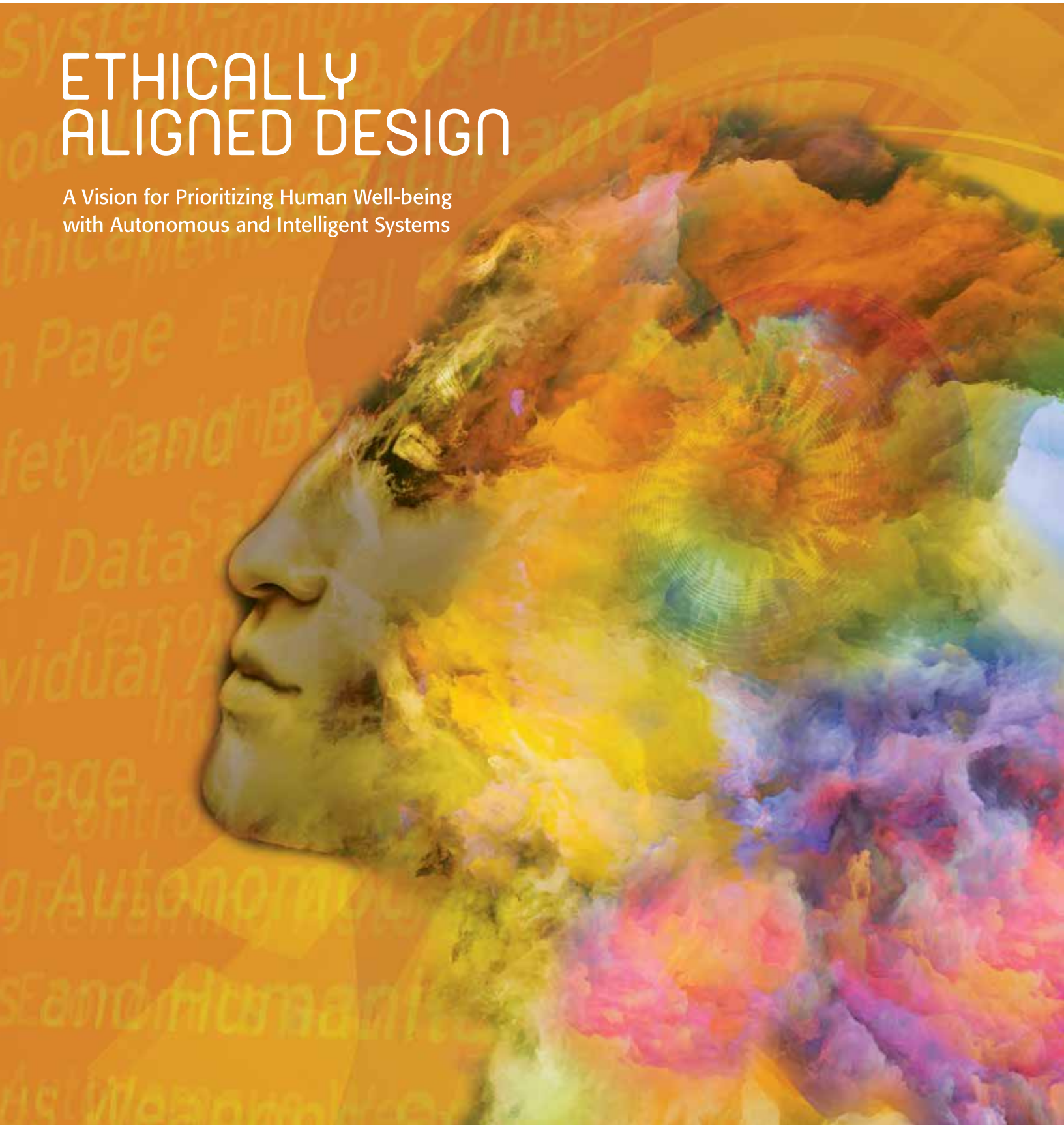
Overview - Version 2

IEEE
Advancing Technology for Humanity

ETHICALLY ALIGNED DESIGN

A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems



Ethically Aligned Design – Version 2 Request for Input

Public comments are invited on the second version of *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (A/IS) that encourages technologists to prioritize ethical considerations in the creation of such systems.

This document has been created by committees of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, ("The IEEE Global Initiative") composed of several hundred participants from six continents, who are thought leaders from academia, industry, civil society, policy and government in the related technical and humanistic disciplines to identify and find consensus on timely issues.

The document's purpose is to:

- Advance a public discussion about how we can establish ethical and social implementations for intelligent and autonomous systems and technologies, aligning them to defined values and ethical principles that prioritize human well-being in a given cultural context.
- Inspire the creation of Standards (IEEE P7000™ series and beyond) and associated certification programs.
- Facilitate the emergence of national and global policies that align with these principles.

By inviting comments for Version 2 of *Ethically Aligned Design*, The IEEE Global Initiative provides the opportunity to bring together multiple voices from the related scientific and engineering communities with the general public to identify and find broad consensus on pressing ethical and social issues and candidate recommendations regarding development and implementations of these technologies.

Input about *Ethically Aligned Design* should be sent by email no later than 7 May 2018 and will be made publicly available at the website of *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems* no later than 4 June 2018. Details on how to submit public comments are available via our Submission Guidelines.

Publicly available comments in response to this request for input will be considered by committees of The IEEE Global Initiative for potential inclusion in the final version of *Ethically Aligned Design* to be released in 2019.

For further information, learn more at the website of The IEEE Global Initiative.

If you're a journalist and would like to know more about The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, please contact the IEEE-SA PR team.

Ethically Aligned Design, Version 2 – Overview

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems is a program of The Institute of Electrical and Electronics Engineers, Inc. (IEEE), the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity with over 420,000 members in more than 160 countries. The IEEE Global Initiative brings together over 250 participants who are thought leaders from academia, industry, civil society and government from six continents in the autonomous and intelligent systems communities to identify and find consensus on timely issues in these fields. The mission of The IEEE Global Initiative is to ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.

In 2016, The IEEE Global Initiative produced *Ethically Aligned Design* (EAD, Version 1), which represents the collective input of this community in the fields of Autonomous and Intelligent Systems (A/IS), ethics, philosophy and policy. EAD inspired, up to now, the creation of eleven IEEE P7000™ Standards Working Groups that reflect the recommendations of The IEEE Global Initiative. The goal of The IEEE Global Initiative is that *Ethically Aligned Design* and the IEEE standards it inspires will provide insights and recommendations that would become a key reference for the work of technologists in the coming years. Below is a synopsis of *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version II*.

Disclaimer: *The recommendations provided in* Ethically Aligned Design *do not represent a position or the views of IEEE but the informed opinions of Committee members providing insights designed to provide expert directional guidance regarding A/IS. In no event shall IEEE or IEEE-SA Industry Connections Activity Members be liable for any errors or omissions, direct or otherwise, however caused, arising in any way out of the use of this work, regardless of whether such damage was foreseeable.*

Ethically Aligned Design, Version 2 – Overview

I. Purpose

Intelligent and autonomous technical systems are specifically designed to reduce human intervention in our day-to-day lives. In so doing, these new fields are raising concerns about their impact on individuals and societies. Current discussions include advocacy for the positive impact, as well as warnings, based on the potential harm to privacy, discrimination, loss of skills, economic impacts, security of critical infrastructure, and the long-term effects on social well-being. Because of their nature, the full benefit of these technologies will be attained only if they are aligned with our defined values and ethical principles. We must therefore establish frameworks to guide and inform dialogue and debate around the non-technical implications of these technologies.

II. Goals

The ethical design, development, and implementation of these technologies should be guided by the following General Principles:

- **Human Rights:** Ensure they do not infringe on internationally recognized human rights
- **Well-being:** Prioritize metrics of well-being in their design and use
- **Accountability:** Ensure that their designers and operators are responsible and accountable
- **Transparency:** Ensure they operate in a transparent manner
- **Awareness of misuse:** Minimize the risks of their misuse

III. Objectives

Personal Data Rights and Individual Access Control

A fundamental need is that people have the right to define access and provide informed consent with respect to the use of their personal digital data. Individuals require mechanisms to help curate their unique identity and personal data in conjunction with policies and practices that make them explicitly aware of consequences resulting from the bundling or resale of their personal information.

Well-being Promoted by Economic Effects

Through affordable and universal access to communications networks and the Internet, intelligent and autonomous technical systems can be made available to and benefit populations anywhere. They can significantly alter institutions and institutional relationships toward more human-centric structures and they can benefit humanitarian and development issues resulting in increased individual and societal well-being.

Legal Frameworks for Accountability

The convergence of intelligent systems and robotics technologies has led to the development of systems with attributes that simulate those of human beings in terms of partial autonomy, ability to perform specific intellectual tasks, and may even have a human physical appearance. The issue of the legal status of complex intelligent and autonomous technical systems thus intertwines with broader legal questions regarding

Ethically Aligned Design, Version 2 – Overview

how to ensure accountability and allocate liability when such systems cause harm. Some examples of general frameworks to consider include the following:

- Intelligent and autonomous technical systems should be subject to the applicable regimes of property law
- Government and industry stakeholders should identify the types of decisions and operations that should never be delegated to such systems and adopt rules and standards that ensure effective human control over those decisions and how to allocate legal responsibility for harm caused by them

Transparency and Individual Rights

Although self-improving algorithms and data analytics can enable the automation of decision-making impacting citizens, legal requirements mandating transparency, participation, and accuracy should include the following objectives:

- Parties, their lawyers, and courts must have reasonable access to all data and information generated and used by such systems employed by governments and other state authorities
- The logic and rules embedded in the system must be available to overseers thereof, if possible, and subject to risk assessments and rigorous testing
- The systems should generate audit trails recording the facts and law supporting decisions and they should be amenable to third-party verification
- The general public should know who is making or supporting ethical decisions of such systems through investment

Policies for Education and Awareness

Effective policy addresses the protection and promotion of safety, privacy, intellectual property rights, human rights, and cybersecurity, as well as the public understanding of the potential impact of intelligent and autonomous technical systems on society. To ensure that they best serve the public interest, policies should:

- Support, promote, and enable internationally recognized legal norms
- Develop workforce expertise in related technologies
- Attain research and development leadership
- Regulate to ensure public safety and responsibility
- Educate the public on societal impacts of related technologies

IV. Foundations

Classical Ethics

By drawing from over two thousand years of classical ethics traditions, The IEEE Global Initiative explores established ethics systems, addressing both scientific and religious approaches, including secular philosophical traditions, to address human morality in the digital age. Through reviewing the philosophical foundations that define autonomy and ontology, The IEEE Global Initiative addresses the alleged potential for autonomous capacity of intelligent technical systems, morality in amoral systems, and asks whether decisions made by amoral systems can have moral consequences.

Ethically Aligned Design, Version 2 – Overview

Well-being Metrics

For extended intelligence and automation to provably advance a specific benefit for humanity, there needs to be clear indicators of that benefit. Common metrics of success include profit, occupational safety, and fiscal health. While important, these metrics fail to encompass the full spectrum of well-being for individuals or society. Psychological, social, and environmental factors matter. Well-being metrics capture such factors, allowing the benefits arising from technological progress to be more comprehensively evaluated, providing opportunities to test for unintended negative consequences that could diminish human well-being. Conversely, these metrics could help identify where intelligent technical systems would increase human well-being, providing new routes to societal and technological innovation.

Embedding Values into Autonomous Systems

If machines engage in human communities as quasi-autonomous agents, then those agents will be expected to follow the community's social and moral norms. Embedding norms in such systems requires a clear delineation of the community in which they are to be deployed. Further, even within a particular community, different types of technical embodiments will demand different sets of norms. The first step is to identify the norms of the specific community in which the systems are to be deployed and, in particular, norms relevant to the kinds of tasks that they are designed to perform.

Methodologies to Guide Ethical Research and Design

To create intelligent technical systems that enhance and extend human well-being and freedom, value-based design methodologies put human advancement at the core of development of technical systems, in concert with the recognition that machines should serve humans and not the other way around. System developers should employ value-based design methodologies in order to create sustainable systems that can be evaluated in terms of both social costs and also advantages that may increase economic value for organizations.

V. Future Technology Concerns

Reframing Autonomous Weapons

Autonomous systems designed to cause physical harm have additional ethical dimensions as compared to both traditional weapons and/or autonomous systems not designed to cause harm. These ethical dimensions include, at least, the following:

- Ensuring meaningful human control of weapons systems
- Designing automated weapons with audit trails to help guarantee accountability and control
- Including adaptive and learning systems that can explain their reasoning and decisions to human operators in transparent and understandable ways
- Training responsible human operators of autonomous systems who are clearly identifiable

Ethically Aligned Design, Version 2 – Overview

- Achieving behavior of autonomous functions that is predictable to their operators
- Ensuring that the creators of these technologies understand the implications of their work
- Developing professional ethical codes to appropriately address the development of autonomous systems intended to cause harm

Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence

Similar to other powerful technologies, the development and use of intelligent and potentially self-improving technical systems involves considerable risk, either because of misuse or poor design. However, according to some theories, as systems approach and surpass AGI, unanticipated or unintended system behavior will become increasingly dangerous and difficult to correct. It is likely that not all AGI-level architectures can be aligned with human interests, and as such, care should be taken to determine how different architectures will perform as they become more capable.

Affective Computing

Affect is a core aspect of intelligence. Drives and emotions such as anger, fear, and joy are often the foundations of actions throughout our life. To ensure that intelligent technical systems will be used to help humanity to the greatest extent possible in all contexts, artifacts participating in or facilitating human society should not cause harm either by amplifying or damping human emotional experience. Even the rudimentary versions of synthetic emotions already deployed in some systems impact how they are perceived by policy makers and the general public.

Mixed Reality

Mixed reality could alter our concepts of identity and reality as these technologies become more common in our work, education, social lives, and commercial transactions. The ability for real-time personalization of this mixed-reality world raises ethical questions concerning the rights of the individual and control over one's multifaceted identity, especially as the technology moves from headsets to more subtle and integrated sensory enhancements.