Executive Summary

To fully benefit from the potential of Artificial Intelligence and Autonomous Systems (AI/AS), we need to go beyond perception and beyond the search for more computational power or solving capabilities.

We need to make sure that these technologies are aligned to humans in terms of our moral values and ethical principles. AI/AS have to behave in a way that is beneficial to people beyond reaching functional goals and addressing technical problems. This will allow for an elevated level of trust between humans and our technology that is needed for a fruitful pervasive use of AI/AS in our daily lives.

Eudaimonia, as elucidated by Aristotle, is a practice that defines human wellbeing as the highest virtue for a society. Translated roughly as “flourishing,” the benefits of eudaimonia begin by conscious contemplation, where ethical considerations help us define how we wish to live.

By aligning the creation of AI/AS with the values of its users and society we can prioritize the increase of human wellbeing as our metric for progress in the algorithmic age.

Executive Summary

Who We Are

The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems ("The IEEE Global Initiative") is a program of The Institute of Electrical and Electronics Engineers, Incorporated ("IEEE"), the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity with over 400,000 members in more than 160 countries.

The IEEE Global Initiative provides the opportunity to bring together multiple voices in the Artificial Intelligence and Autonomous Systems communities to identify and find consensus on timely issues.

IEEE will make *Ethically Aligned Design* (EAD) available under the Creative Commons Attribution-Non-Commercial 3.0 United States License.

Subject to the terms of that license, organizations or individuals can adopt aspects of this work at their discretion at any time. It is also expected that EAD content and subject matter will be selected for submission into formal IEEE processes, including for standards development.

The IEEE Global Initiative and EAD contribute to a broader effort being launched at IEEE to foster open, broad and inclusive conversation about ethics in technology, known as the IEEE TechEthics™ program.

The Mission of The IEEE Global Initiative

To ensure every technologist is *educated*, *trained*, and *empowered* to prioritize ethical considerations in the design and development of autonomous and intelligent systems.

By "technologist", we mean anyone involved in the research, design, manufacture or messaging around AI/AS including universities, organizations, and corporations making these technologies a reality for society.

This document represents the collective input of over one hundred global thought leaders in the fields of Artificial Intelligence, law and ethics, philosophy, and policy from the realms of academia, science, and the government and corporate sectors. Our goal is that *Ethically Aligned Design* will provide insights and recommendations from these peers that provide a key reference for the work of AI/AS technologists in the coming years. To achieve this goal, in the current version of *Ethically Aligned Design* (EAD v1), we identify Issues and Candidate Recommendations in fields comprising Artificial Intelligence and Autonomous Systems.

A second goal of The IEEE Global Initiative is to provide recommendations for IEEE Standards based on *Ethically Aligned Design*. IEEE P7000™ *– Model Process for Addressing Ethical Concerns During System Design* was the first IEEE Standard Project (approved and in development) inspired by The Initiative. Two further Standards Projects, IEEE P7001TM – Transparency of Autonomous Systems and IEEE P7002TM – Data Privacy Process, have been approved, demonstrating The Initiative's pragmatic influence on issues of AI/AS ethics.

Structure and Content of the Document

Ethically Aligned Design includes eight sections, each addressing a specific topic related to AI/AS that has been discussed at length by a specific committee of The IEEE Global Initiative. Issues and candidate recommendations pertaining to these topics are listed in each committee section. Below is a summary of the committees and the issues covered in their sections:

1 | General Principles

The General Principles Committee has articulated high-level ethical concerns applying to all types of AI/AS that:

1. Embody the highest ideals of human rights.
2. Prioritize the maximum benefit to humanity and the natural environment.
3. Mitigate risks and negative impacts as AI/AS evolve as socio-technical systems.

It is the Committee's intention that the Principles, Issues, and Candidate Recommendations they have identified will eventually serve to underpin and scaffold future norms and standards within a new framework of ethical governance for AI/AS design.

Issues:

- How can we ensure that AI/AS do not infringe human rights? (Framing the Principle of Human Rights)
- How can we assure that AI/AS are accountable? (Framing the Principle of Responsibility)
- How can we ensure that AI/AS are transparent? (Framing the Principle of Transparency)
- How can we extend the benefits and minimize the risks of AI/AS technology being misused? (Framing the Principle of Education and Awareness)

2 | Embedding Values into Autonomous Intelligence Systems

In order to develop successful autonomous intelligent systems (AIS) that will benefit society, it is crucial for the technical community to understand and be able to embed relevant human norms or values into their systems. The *Embedding Values into Autonomous Intelligence Systems Committee* has taken on the broader objective of embedding values into AIS as a three-pronged approach by helping designers:

1. Identify the norms and values of a specific community affected by AIS;

2. Implement the norms and values of that community within AIS; and,

3. Evaluate the alignment and compatibility of those norms and values between the humans and AIS within that community.

Issues:

- Values to be embedded in AIS are not universal, but rather largely specific to user communities and tasks.

- Moral overload: AIS are usually subject to a multiplicity of norms and values that may conflict with each other.

- AIS can have built-in data or algorithmic biases that disadvantage members of certain groups.

- Once the relevant sets of norms (of AIS's specific role in a specific community) have been identified, it is not clear how such norms should be built into a computational architecture.

- Norms implemented in AIS must be compatible with the norms in the relevant community.

- Achieving a correct level of trust between humans and AIS.

- Third-party evaluation of AIS's value alignment.

3 | Methodologies To Guide Ethical Research and Design

The modern AI/AS organization should ensure that human wellbeing, empowerment, and freedom are at the core of AI/AS development. To create machines that can achieve these ambitious goals the Methodologies to Guide Ethical Research and Design Committee has framed issues and candidate recommendations to ensure that human values, like human rights as defined in the Universal Declaration of Human Rights, are engendered by their system design methodologies. Values-aligned design methodologies should become an essential focus for AI/AS organizations, geared to human advancement based on ethical guidelines. Machines should serve humans and not the other way around. This ethically sound approach will ensure that an equal balance is struck between preserving the economic and the social affordances of AI, for both business and society.

Issues:

- Ethics is not part of degree programs.

- We need models for interdisciplinary and intercultural education to account for the distinct issues of AI/AS.

- The need to differentiate culturally distinctive values embedded in AI design.

- Lack of value-based ethical culture and practices for industry.

- Lack of values-aware leadership.

- Lack of empowerment to raise ethical concerns.
- Lack of ownership or responsibility from tech community.
- Need to include stakeholders for best context of AI/AS.
- Poor documentation hinders ethical design.
- Inconsistent or lacking oversight for algorithms.
- Lack of an independent review organization.
- Use of black-box components.

4 | **Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI)**

Future highly capable AI systems (sometimes referred to as artificial general intelligence or AGI) may have a transformative effect on the world on the scale of the agricultural or industrial revolutions, which could bring about unprecedented levels of global prosperity. The Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI) Committee has provided multiple issues and candidate recommendations to help ensure this transformation will be a positive one via the concerted effort by the AI community to shape it that way.

Issues:

- As AI systems become more capable—as measured by the ability to optimize more complex objective functions with greater autonomy across a wider variety of domains—unanticipated or unintended behavior becomes increasingly dangerous.
- Retrofitting safety into future, more generally capable, AI systems may be difficult.
- Researchers and developers will confront a progressively more complex set of ethical and technical safety issues in the development and deployment of increasingly autonomous and capable AI systems.
- Future AI systems may have the capacity to impact the world on the scale of the agricultural or industrial revolutions.

5 | **Personal Data and Individual Access Control**

A key ethical dilemma regarding personal information is *data asymmetry*. To address this asymmetry the Personal Data and Individual Access Control Committee has elucidated issues and candidate recommendations demonstrating the fundamental need for people to *define*, *access*, and *manage* their personal data as curators of their unique identity. The Committee recognizes there are no perfect solutions, and that any digital tool can be hacked. Nonetheless they recommend the enablement of a data environment where people control their sense of self and have provided examples of tools and evolved practices that could eradicate data asymmetry for a positive future.

Issues:

- How can an individual define and organize his/her personal data in the algorithmic era?
- What is the definition and scope of personally identifiable information?
- What is the definition of control regarding personal data?
- How can we redefine data access to honor the individual?
- How can we redefine consent regarding personal data so it honors the individual?
- Data that appears trivial to share can be used to make inferences that an individual would not wish to share.
- How can data handlers ensure the consequences (positive and negative) of accessing and collecting data are explicit to an individual in order to give truly informed consent?
- Could a person have a personalized AI or algorithmic guardian?

6 | Reframing Autonomous Weapons Systems

Autonomous systems that are designed to cause physical harm have additional ethical ramifications as compared to both traditional weapons and autonomous systems that aren't designed to cause harm. Professional ethics about these can and should have a higher standard covering a broader array of concerns. Broadly, the Reframing Autonomous Weapons Systems Committee recommends that technical organizations accept that meaningful human control of weapons systems is beneficial to society, that audit trails guaranteeing accountability ensure such control, that those creating these technologies understand the implications of their work, and that professional ethical codes appropriately address works that are intended to cause harm.

Issues:

- Professional organization codes of conduct often have significant loopholes, whereby they overlook holding members' works, the artifacts and agents they create, to the same values and standards that the members themselves are held to, to the extent that those works can be.
- Confusions about definitions regarding important concepts in artificial intelligence, autonomous systems, and autonomous weapons systems (AWS) stymie more substantive discussions about crucial issues.
- AWS are by default amenable to covert and non-attributable use.

- There are multiple ways in which accountability for AWS's actions can be compromised.
- AWS might not be predictable (depending upon its design and operational use). Learning systems compound the problem of predictable use.
- Legitimizing AWS development sets precedents that are geopolitically dangerous in the medium-term.
- Exclusion of human oversight from the battlespace can too easily lead to inadvertent violation of human rights and inadvertent escalation of tensions.
- The variety of direct and indirect customers of AWS will lead to a complex and troubling landscape of proliferation and abuse.
- By default, the type of automation in AWS encourage rapid escalation of conflicts.
- There are no standards for design assurance verification of AWS.
- Understanding the ethical boundaries of work on AWS and semi-autonomous weapons systems can be confusing.

7 | Economics/Humanitarian Issues

Technologies, methodologies, and systems that aim to reduce human intervention in our day-to-day lives are evolving at a rapid pace and are poised to transform the lives of individuals in multiple ways. The aim of the Economics/Humanitarian Issues Committee is to identify the key drivers shaping the human-technology global ecosystem and address economic and humanitarian ramifications, and to suggest key opportunities for solutions that could be implemented by unlocking critical choke points of tension. The goal of the Committee's recommendations is to suggest a pragmatic direction related to these central concerns in the relationship of humans, their institutions and emerging information-driven technologies, to facilitate interdisciplinary, cross-sector dialog that can be more fully informed by expert, directional, and peer-guided thinking regarding these issues.

Issues:

- Misinterpretation of AI/AS in media is confusing to the public.
- Automation is not typically viewed only within market contexts.
- The complexities of employment are being neglected regarding robotics/AI.
- Technological change is happening too fast for existing methods of (re)training the workforce.
- Any AI policy may slow innovation.

- AI and autonomous technologies are not equally available worldwide.
- There is a lack of access and understanding regarding personal information.
- An increase of active representation of developing nations in The IEEE Global Initiative is needed.
- The advent of AI and autonomous systems can exacerbate the economic and power-structure differences between and within developed and developing nations.

8 | **Law**

The early development of AI/AS has given rise to many complex ethical problems. These ethical issues almost always directly translate into concrete legal challenges—or they give rise to difficult collateral legal problems. The Law Committee feels there is much work for lawyers in this field that, thus far, has attracted very few practitioners and academics despite being an area of pressing need. Lawyers need to be part of discussions on regulation, governance, domestic and international legislation in these areas so the huge benefits available to humanity and our planet from AI/AS are thoughtfully stewarded for the future.

Issues:

- How can we improve the accountability and verifiability in autonomous and intelligent systems?
- How can we ensure that AI is transparent and respects individual rights? For example, international, national, and local governments are using AI which impinges on the rights of their citizens who should be able to trust the government, and thus the AI, to protect their rights.
- How can AI systems be designed to guarantee legal accountability for harms caused by these systems?
- How can autonomous and intelligent systems be designed and deployed in a manner that respects the integrity of personal data?

Our New Committees and their current work are described at the end of *Ethically Aligned Design*.

How the Document was Prepared

This document was prepared using an open, collaborative and consensus building approach, following the processes of the Industry Connections program, a program of the IEEE Standards Association. Industry Connections facilitates collaboration among organizations and individuals as they hone and refine their thinking on emerging technology issues, helping to incubate potential new standards activities and standards related products and services.

How to Cite Ethically Aligned Design

Please cite Version 1 of *Ethically Aligned Design* in the following manner:

The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. *Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems*, Version 1. IEEE, 2016. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

Our Appreciation

We wish to express our appreciation for the organizations who have recently contributed research and insights helping to increase awareness around ethical issues and AI/AS, including (but not limited to): AI Now (White House/New York University); One Hundred Year Study on Artificial Intelligence (Stanford University); Preparing for The Future of Artificial Intelligence (U.S. White House/NSTC); The National Artificial Intelligence Research and Development Strategic Plan (U.S. White House/NSTC); Robotics and Artificial Intelligence (U.K. House of Commons Science and Technology Committee); Robots and Robotic Devices – Guide to the Ethical Design and Application of Robots and Robotic Systems (British Standards Institute); Japan's Basic Rules for AI Research; Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics (European Parliament); Éthique de la recherche en robotique (CERNA); Charta der Digitalen Grundrechte der Europäischen Union (Charter of the Digital Fundamental Rights of the European Union); and, Research Priorities for Robust and Beneficial Artificial Intelligence (Future of Life Institute).

We also wish to express our appreciation for the following organizations regarding their seminal efforts regarding AI/AS Ethics, including (but not limited to): The Association for the Advancement of Artificial Intelligence and their formative work on AI Ethics; European Association for Artificial Intelligence; ACM Special Interest Group on Artificial Intelligence; The IEEE Robot and Automation Society Committee on Robot Ethics; The IEEE Society on Social Implications of Technology; The Leverhulme Centre for the Future of Intelligence; Allen Institute for Artificial Intelligence; OpenAI; Machine Intelligence Research Institute; Centre for The Study of Existential Risk; AI-Austin and, Partnership on AI to Benefit People and Society.

We would also like to acknowledge the contribution of Eileen M. Lach, the General Counsel and Chief Compliance Officer of IEEE, who has reviewed this document in its entirety and affirms the importance of the contribution of The IEEE Global Initiative to the fields of AI/AS ethics.

Executive Summary

Disclaimers

Ethically Aligned Design is not a code of conduct or a professional code of ethics. Engineers and technologists have well-established codes, and we wish to respectfully recognize the formative precedents surrounding issues of ethics and safety and the professional values these Codes represent. These Codes provide the broad framework for the more focused domain of AI/AS addressed in this document, and it is our hope that the inclusive, consensus- building process around its design will contribute unique value to technologists and society as a whole.

This document is also not a position, or policy statement, or formal report. It is intended to be a working reference tool created in an inclusive process by those in the AI/AS Community prioritizing ethical considerations in their work.

A Note on Affiliations Regarding Members of The Initiative

The language and views expressed in *Ethically Aligned Design* reflect the individuals who created content for each section of this document. The language and views expressed in this document do not necessarily reflect the Universities or Organizations to which these individuals belong, and should in no way be considered any form of endorsement, implied or otherwise, from these institutions.

This is a first version of *Ethically Aligned Design*. Where individuals are listed in a Committee it indicates only that they are Members of that Committee. Committee Members may not have achieved final consensus on content in this document because of its versioning format and the consensus-building process of The IEEE Global Initiative for Ethical Consideration in Artificial Intelligence and Autonomous Systems. Content listed by Members in this or future versions is not an endorsement, implied or otherwise, until formally stated as such.

A Note Regarding Candidate Recommendations in this Document

Ethically Aligned Design is being created via multiple versions that are being iterated over the course of two to three years. The IEEE Global Initiative is following a specific consensus-building process where members contributing content are proposing candidate recommendations so as not to imply these are final recommendations at this time.

Our Membership

Although The IEEE Global Initiative currently has more than one hundred experts from all but one continent involved in our work, most of us come from North America and Europe. We are aware we need to expand our cultural horizons and have more people involved from around the world as we continue to grow our document and our efforts. We are eager for these new voices and perspectives to join our work.

Trademarks and Disclaimers

IEEE believes in good faith that the information in this publication is accurate as of its publication date; such information is subject to change without notice. IEEE is not responsible for any inadvertent errors.

The Institute of Electrical and Electronics
Engineers, Incorporated
3 Park Avenue, New York, NY 10016-5997, USA

Executive Summary

All rights reserved. Published Month 20xx.
Printed in the United States of America.

IEEE is a registered trademark in the U. S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated.

PDF: ISBN 978-0-7381-xxxx-x STDVxxxxx

Print: ISBN 978-0-7381-xxxx-x STDPDVxxxxx

IEEE prohibits discrimination, harassment, and bullying. For more information, visit http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html.

To order IEEE Press Publications, call 1-800-678-IEEE.

Find IEEE standards and standards-related product listings at: http://standards.ieee.org

Notice and Disclaimer of Liability Concerning the Use of IEEE-SA Industry Connections Documents

This IEEE Standards Association ("IEEE-SA") Industry Connections publication ("Work") is not a consensus standard document. Specifically, this document is NOT AN IEEE STANDARD. Information contained in this Work has been created by, or obtained from, sources believed to be reliable, and reviewed by members of the IEEE-SA Industry Connections activity that produced this Work. IEEE and the IEEE-SA Industry Connections activity members expressly disclaim all warranties (express, implied, and statutory) related to this Work, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; quality, accuracy, effectiveness, currency, or completeness of the Work or content within the Work. In addition, IEEE and the IEEE-SA Industry Connections activity members disclaim any and all conditions relating to: results; and workmanlike effort. This IEEE-SA Industry Connections document is supplied "AS IS" and "WITH ALL FAULTS."

Although the IEEE-SA Industry Connections activity members who have created this Work believe that the information and guidance given in this Work serve as an enhancement to users, all persons must rely upon their own skill and judgment when making use of it. IN NO EVENT SHALL IEEE OR IEEE-SA INDUSTRY CONNECTIONS ACTIVITY MEMBERS BE LIABLE FOR ANY ERRORS OR OMISSIONS OR DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Further, information contained in this Work may be protected by intellectual property rights held by third parties or organizations, and the use of this information may require the user to negotiate with any such rights holders in order to legally acquire the rights to do so, and such rights holders may refuse to grant such rights. Attention is also called to the possibility that implementation of any or all of this Work may require use of subject matter covered by patent rights. By publication of this Work, no position is taken by IEEE with respect to the existence or validity of any patent rights in connection therewith. IEEE is not responsible for identifying patent rights for which a license may be required, or for conducting inquiries into the legal validity or scope of patents claims. Users are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. No commitment to grant licenses under patent rights on a reasonable or non-discriminatory basis has been sought or received from any rights holder. The policies and procedures under which this document was created can be viewed at http://standards.ieee.org/about/sasb/iccom/.

This Work is published with the understanding that IEEE and the IEEE-SA Industry Connections activity members are supplying information through this Work, not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought. IEEE is not responsible for the statements and opinions advanced in this Work.