OVERVIEW

IEEE
Advancing Technology
for Humanity

ETHICALLY ALIGNED DESIGN

A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems

ETHICALLY ALIGNED DESIGN–VERSION ONE

REQUEST FOR INPUT

Public comments are invited on *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems* (AI/AS) that encourages technologists to prioritize ethical considerations in the creation of autonomous and intelligent technologies. This document has been created by committees of *The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems*, comprised of over one hundred global thought leaders and experts in artificial intelligence, ethics, and related issues.

The following Overview is an introduction to Ethically Aligned Design, a document driven by The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. Download the complete document.

The document's purpose is to advance a public discussion of how these intelligent and autonomous technologies can be aligned to moral values and ethical principles that prioritize human wellbeing.

By inviting comments for Version One of *Ethically Aligned Design*, The IEEE Global Initiative provides the opportunity to bring together multiple voices from the Artificial Intelligence and Autonomous Systems (AI/AS) communities with the general public to identify and find broad consensus on pressing ethical issues and candidate recommendations regarding these technologies.

Input about *Ethically Aligned Design* should be sent by e-mail no later than 6 March 2017 and will be made publicly available at The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems no later than 10 April 2017. Details on how to submit public comments are available via the Submission Guidelines.

New and existing committees contributing to an updated version of *Ethically Aligned Design* will be featured at The IEEE Global Initiative's face-to-face meeting at The Robert S. Strauss Center at The University of Texas at Austin to be held 5-6 June 2017. Publicly available comments in response to this request for input will be considered by committees and participants of the meeting for potential inclusion in Version Two of *Ethically Aligned Design* to be released in the fall of 2017.

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

If you're a journalist and would like to know more about The IEEE Global Initiative for Ethically Aligned Design, please contact the IEEE-SA PR team.

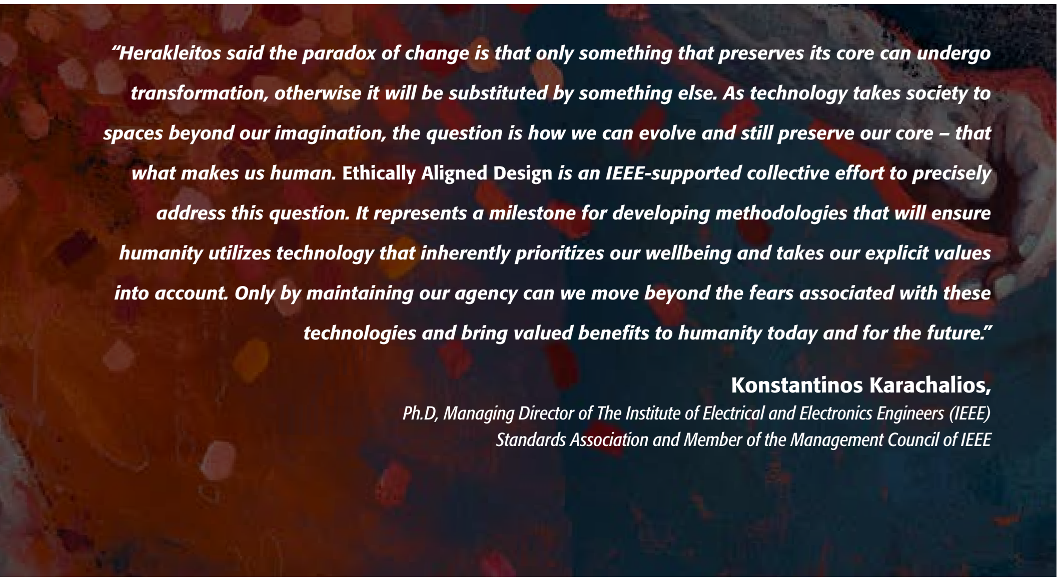
INTRODUCTION

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.

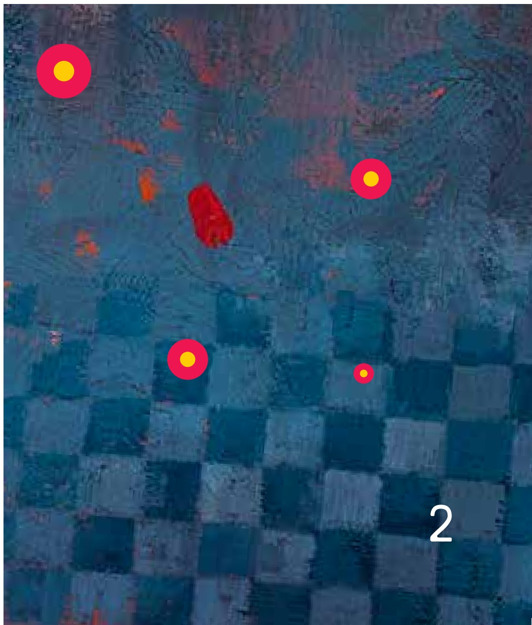
Eudaimonia, as elucidated by Aristotle, is a practice that defines human wellbeing as the highest virtue for a society. Translated roughly as "flourishing," the benefits of eudaimonia begin by conscious contemplation, where ethical considerations help us define how we wish to live.

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.



"Herakleitos said the paradox of change is that only something that preserves its core can undergo transformation, otherwise it will be substituted by something else. As technology takes society to spaces beyond our imagination, the question is how we can evolve and still preserve our core – that what makes us human.* Ethically Aligned Design *is an IEEE-supported collective effort to precisely address this question. It represents a milestone for developing methodologies that will ensure humanity utilizes technology that inherently prioritizes our wellbeing and takes our explicit values into account. Only by maintaining our agency can we move beyond the fears associated with these technologies and bring valued benefits to humanity today and for the future."

Konstantinos Karachalios,
Ph.D, Managing Director of The Institute of Electrical and Electronics Engineers (IEEE)
Standards Association and Member of the Management Council of IEEE



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.

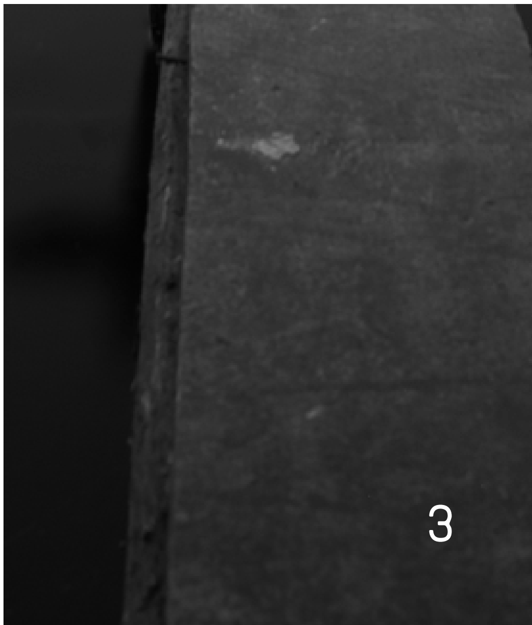
The IEEE Global Initiative and the EAD contribute to a broader effort being launched at IEEE to foster open, broad and inclusive conversation about ethics in technology, known as the IEEE TechEthics™ program.

"Ethically Aligned Design ***and the work of our Global Initiative is focused on empowering technologists to prioritize ethical considerations in the creation of Artificial Intelligence and Autonomous Systems. Rather than assume a machine or system will de facto provide positive benefits, we must determine and align with the values of society before its implementation."***

Raja Chatila,
(Initiative Chair) CNRS-Sorbonne UPMC Institute of Intelligent Systems and Robotics, Paris, France;
Member of the French Commission on the Ethics of Digital Sciences and Technologies CERNA;
Past President of IEEE Robotics and Automation Society

"As a society, we cannot move forward in a spirit of fear around the creation of Artificial Intelligence and Autonomous Systems. By ensuring ethical methodologies become industry standard in the creation of these technologies we'll shift from a spirit of paranoia to pragmatism and redefine innovation around which machines or systems best honor the values of its users."

Kay Firth-Butterfield,
(Initiative Vice-Chair) Executive Director, AI Austin



THE MISSION OF THE INITIATIVE

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

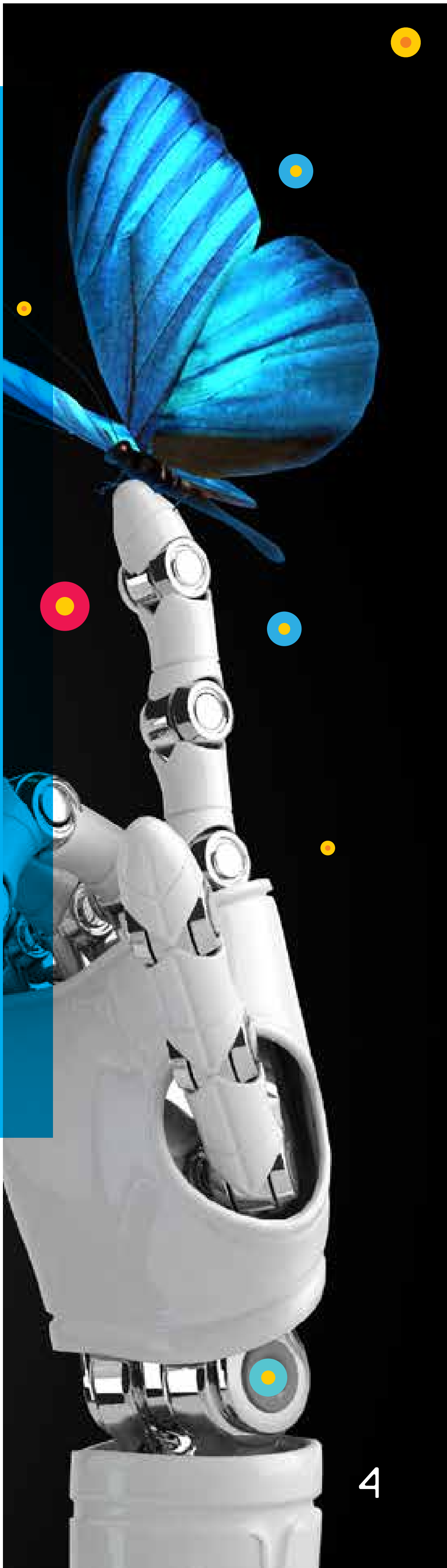
By “technologist”, we mean anyone involved in the research, design, manufacture or messaging around AI/AS including universities, organizations, and corporations making these technologies a reality for society.

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 P7001™ – Transparency of Autonomous Systems and IEEE P7002™ – Data Privacy Process, have been approved, demonstrating The Initiative’s pragmatic influence on issues of AI/AS ethics.

Ethically Aligned Design **includes eight sections, each tackling a specific topic related to AI/AS that has been discussed at length by a specific committee of The IEEE Global Initiative. You can learn more about their work (along with our new Committees) by reading the descriptions of each Committee with quotes from their Chairs in the pages that follow.**

> ***“How will machines know what we value if we don’t know ourselves? Ethics and values-driven design provide tools for introspection technologists should prioritize as we build the machines and systems guiding our lives for the future. We can’t positively increase human wellbeing if we don’t take the time to identify our collective values before creating technology we know will align with those ideals”***
>
> **John C. Havens,**
> *Executive Director of The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, author, Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*



1 General Principles

The General Principles Committee seeks to articulate high-level ethical concerns that apply to all types of artificial intelligence and autonomous systems that:

1) Embody the highest ideals of human rights that honor their inherent dignity and worth.

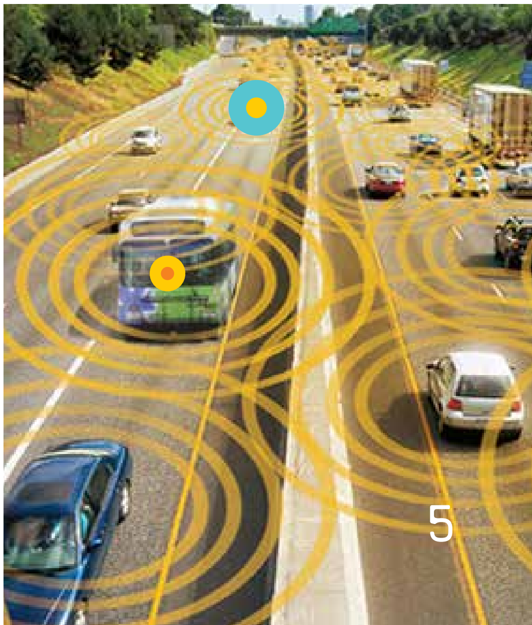
2) Prioritize the maximum benefit to humanity and the natural environment.

3) Mitigate risks and negative impacts as AI/AS evolve as socio-technical systems.

It is our intention that by identifying issues and candidate recommendations regarding these principles they will eventually serve to underpin and scaffold future norms and standards within a new framework of ethical governance.

> ***"For autonomous and intelligent systems to be trusted, and hence bring the greatest benefit, they must be designed and operated ethically. It is vital therefore that we build such systems on a strong foundation of ethical principles."***
>
> **Alan Winfield,**
> *(Co-Chair) Professor, Bristol Robotics Laboratory, University of the West of England; Visiting Professor, University of York*

> ***"The Principles that best honor human dignity should be mirrored in the ethical considerations we utilize when creating future technologies. Artificial Intelligence and Autonomous Systems should prioritize maximum benefit for humanity to ensure society flourishes long into the future."***
>
> **Kay Firth-Butterfield,**
> *(Co-Chair) Executive Director, AI Austin*



2 EMBEDDING VALUES INTO AUTONOMOUS INTELLIGENT SYSTEMS

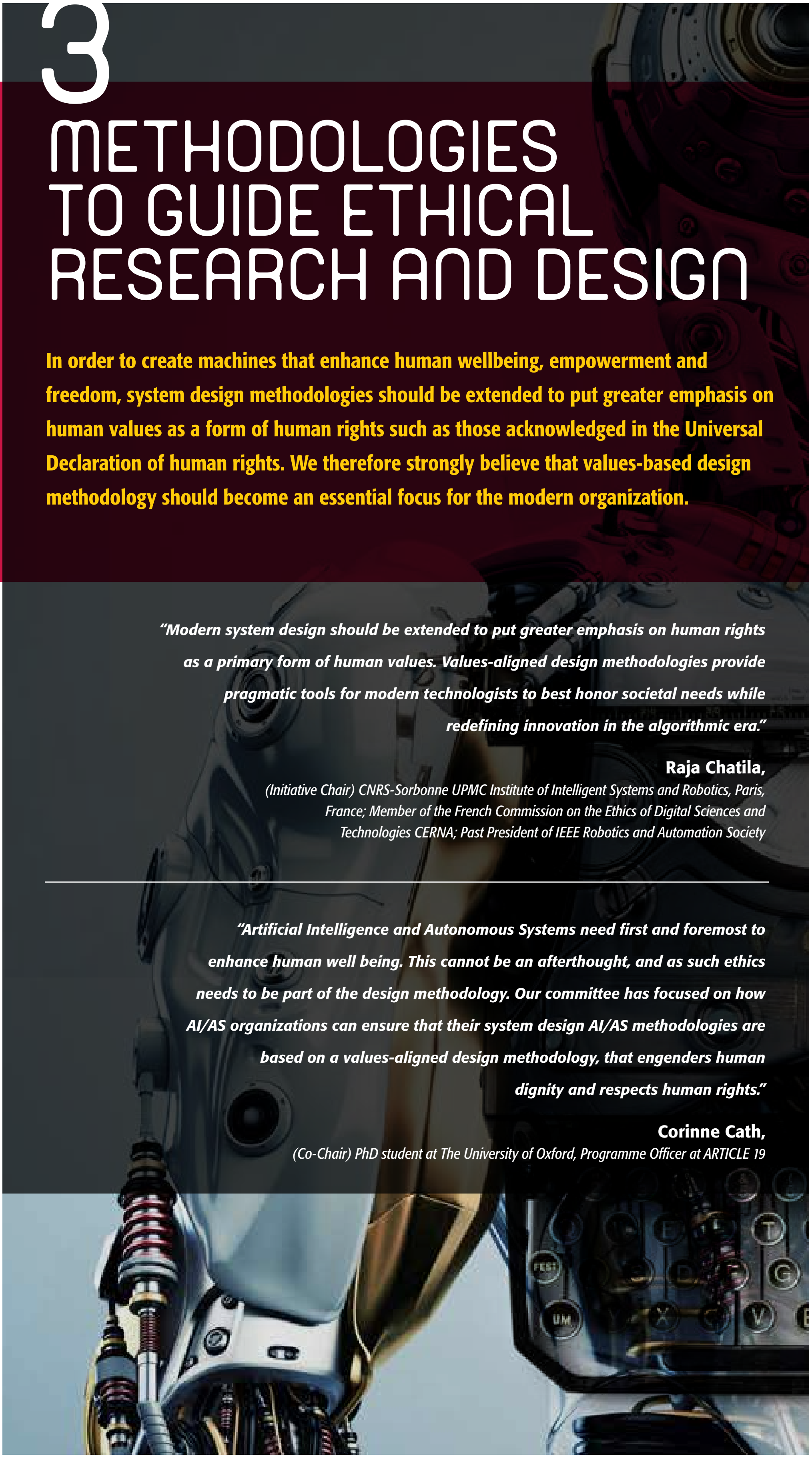
In order to develop successful Autonomous Intelligent Systems (AIS) that will benefit our society, it is crucial for the technical community to understand and be able to embed relevant human norms or values into their systems.

Our Committee has taken on the broader objective of embedding values into AIS as a three-pronged approach, that is to help designers:

1) **Identify** the norms and values of a specific community affected by an AIS;
2) **Implement** the norms and values of that community within the AIS; and,
3) **Evaluate** the alignment and compatibility of those norms and values between the humans and the AIS within that community.

> ***"The alignment of values between a system and its user is of critical importance to ensure Artificial Intelligence and Autonomous Systems increase human wellbeing while optimizing innovation."***
>
> **AJung Moon,**
> *(Co-Chair) Co-founder of the Open Roboethics initiative, and PhD Candidate and Vanier Scholar at the Department of Mechanical Engineering, University of British Columbia*

> ***"In both autonomous systems and human-machine environments, it is essential that the AI system functions according to the correct moral values, social norms, and professional codes. This will allow the building of the correct level of trust between humans and AI. Companies prioritizing these issues, besides the technical capabilities to achieve the specified goals, will have a market advantage over competitors who ignore their critical importance."***
>
> **Francesca Rossi,**
> *(Co-Chair), IBM Research, Yorktown Heights, NY, and full Professor of computer science at the University of Padova, Italy.*



3 METHODOLOGIES TO GUIDE ETHICAL RESEARCH AND DESIGN

In order to create machines that enhance human wellbeing, empowerment and freedom, system design methodologies should be extended to put greater emphasis on human values as a form of human rights such as those acknowledged in the Universal Declaration of human rights. We therefore strongly believe that values-based design methodology should become an essential focus for the modern organization.

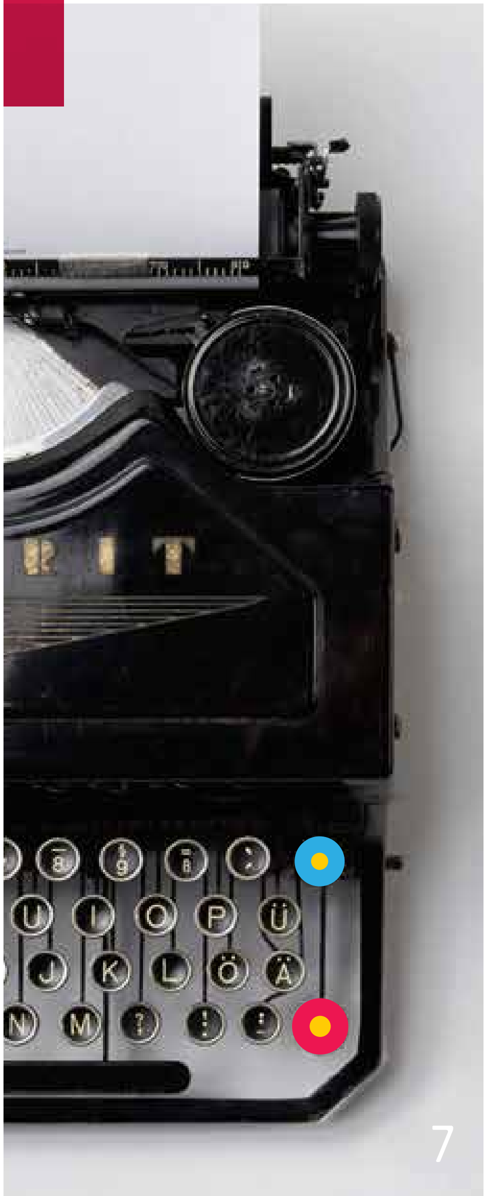
"Modern system design should be extended to put greater emphasis on human rights as a primary form of human values. Values-aligned design methodologies provide pragmatic tools for modern technologists to best honor societal needs while redefining innovation in the algorithmic era."

Raja Chatila,
(Initiative Chair) CNRS-Sorbonne UPMC Institute of Intelligent Systems and Robotics, Paris, France; Member of the French Commission on the Ethics of Digital Sciences and Technologies CERNA; Past President of IEEE Robotics and Automation Society

---

"Artificial Intelligence and Autonomous Systems need first and foremost to enhance human well being. This cannot be an afterthought, and as such ethics needs to be part of the design methodology. Our committee has focused on how AI/AS organizations can ensure that their system design AI/AS methodologies are based on a values-aligned design methodology, that engenders human dignity and respects human rights."

Corinne Cath,
(Co-Chair) PhD student at The University of Oxford, Programme Officer at ARTICLE 19



4

SAFETY & BENEFICENCE OF ARTIFICIAL GENERAL INTELLIGENCE (AGI) & 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 revolution, which could bring about unprecedented levels of global prosperity. It is by no means guaranteed however that this transformation will be a positive one without a concerted effort by the AI community to shape it that way.

"The AI community needs to encourage and promote the sharing and use of safety related research and tools, and generally bring consideration of beneficence in to their work more."

Richard Mallah,
(Co Chair) – Director of AI Projects,
Future of Life Institute

"As AI systems become more useful and capabilities increase, unintended behaviors and accidents will pose correspondingly greater risks. It's essential that the AI community adopt some best practices from computer security, where systems and their safety/security measures are subjected to highly rigorous assessments before seeing wide adoption."

Malo Bourgon,
(Co-Chair) – COO, Machine Intelligence Research Institute

5 PERSONAL DATA AND INDIVIDUAL ACCESS CONTROL

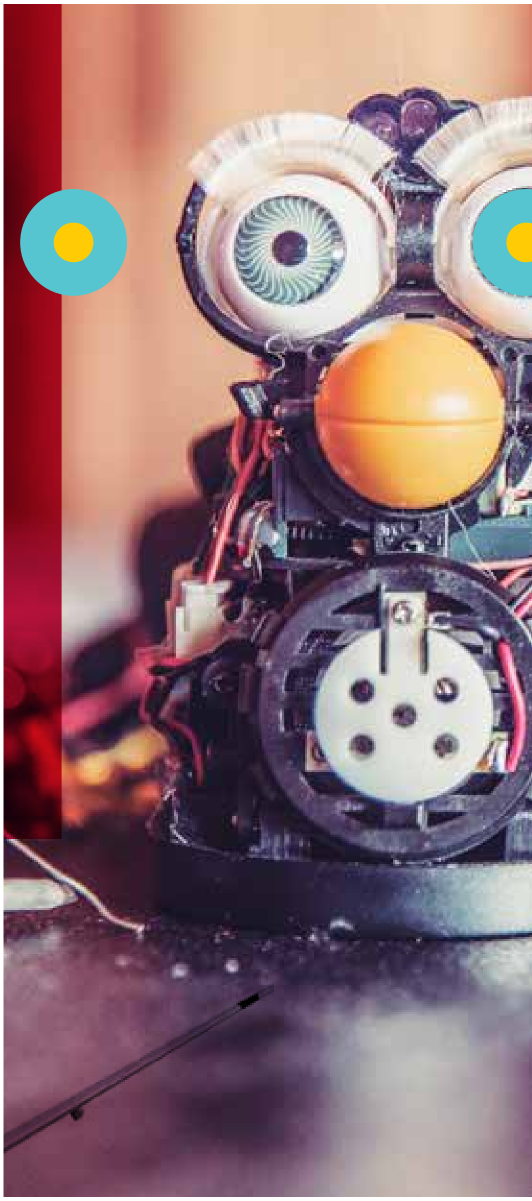
A key ethical dilemma regarding personal information is data asymmetry. To address this asymmetry there is a fundamental need for people to define, access, and manage their personal data as curators of their unique identity. We realize there are no perfect solutions, and that any digital tool can be hacked. But we need to enable a data environment where people control their sense of self. Our goal is to envision the tools and evolved practices that will eradicate data asymmetry to project a positive image of our future.

> ***"Personal Data forms the bedrock of the algorithmic economy. Prioritizing ethical considerations regarding the use of this data by autonomous and intelligent technologies means we'll help individuals gain clarity around their digital assets while improving the quality of information provided to the systems we're building to best guide our future."***
>
> **Michelle Dennedy,**
> *(Co-Chair) Vice President, Chief Privacy Officer, Cisco;*
> *Author, The Privacy Engineer's Manifesto: Getting from Policy to Code to QA to Value*

> ***"Along with personalization methodologies that track our actions, we need tools to manage the personal data reflecting our intentions and subjective identity. Ethical considerations for AI/AS must account for and align with these values to best increase human wellbeing."***
>
> **John C. Havens,**
> *(Co-Chair) Executive Director, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, author, Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*

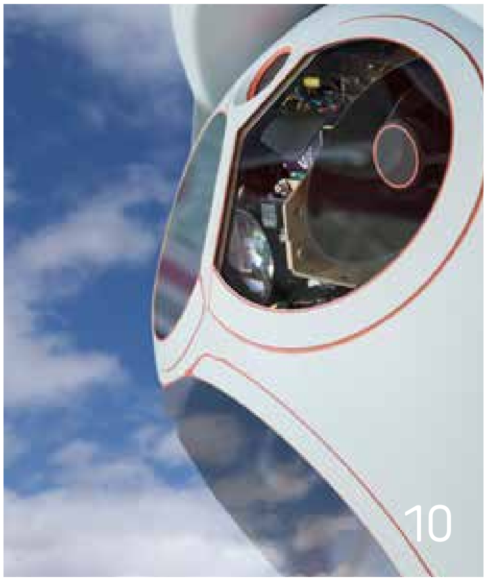
6
REFRAMING AUTONOMOUS WEAPONS SYSTEMS

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, we recommend 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 implications of their work, and that professional ethical codes appropriately address works that are intended to cause harm.



"Ethical considerations and codes of Ethics designed to guide technologists creating autonomous weapons systems need to prioritize meaningful human control for the systems they create."

Richard Mallah,
(Chair) – Director of AI Projects, Future of Life Institute.



7 ECONOMICS/ HUMANITARIAN ISSUES

Technologies, methodologies, and systems that aim at reducing 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 our multi-stakeholder committee is to identify the key drivers shaping the human-technology global ecosystem and address economical and humanitarian ramifications, and to suggest key opportunities for solutions that could be implemented by unlocking critical choke points of tension. The goal of our 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 dialogue that can be more fully informed by expert, directional, and peer-guided thinking regarding these issues.

> *"Artificial Intelligence and Autonomous Systems need to prioritize ethical considerations in their design to ensure the equal distribution of their benefits while reducing harm to society. Otherwise they will not be designed for the increase of wellbeing for all of humanity but simply for those individuals who are privileged to reap their benefits"*
>
> **Raj Madhavan,**
> *(Chair) Founder & CEO of Humanitarian Robotics Technologies, LLC, Maryland, U.S.A.*

8
LAW

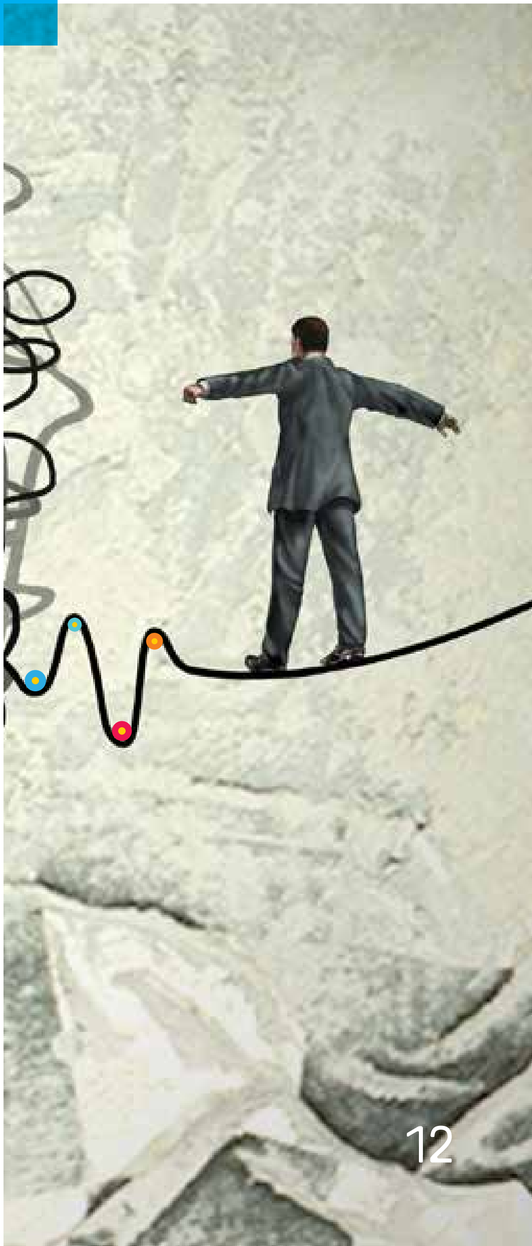
The early development of Artificial Intelligence and Autonomous Systems (AI/AS) has given rise to many complex ethical problems. These ethical issues almost always directly translate into concrete legal challenges – or they give rise to difficult collateral legal problems. There is much to do for lawyers in this field that thus far has attracted very few practitioners and academics despite being an area of pressing need. Lawyers should be part of discussions on regulation, governance, domestic and international legislation in these areas and we welcome this opportunity to ensure that the huge benefits available to humanity and our planet from AI/AS are thoughtfully stewarded for the future.

"It is essential that the laws created to guide AI/AS are built to best honor the community and societal values of the communities in which they're developed."

Derek Jinks,
(Co-Chair) University of Texas Law School; Consortium on Law and Ethics of Artificial Intelligence and Robotics, Strauss Center, University of Texas

"Increasing human wellbeing means creating and adapting our laws to mirror the values we want to develop in ourselves, society, and the machines we build in the future."

Kay Firth-Butterfield,
(Co-Chair) Executive Director, AI Austin



NEW COMMITTEES[1]

Classical Ethics in Information & Communication Technologies

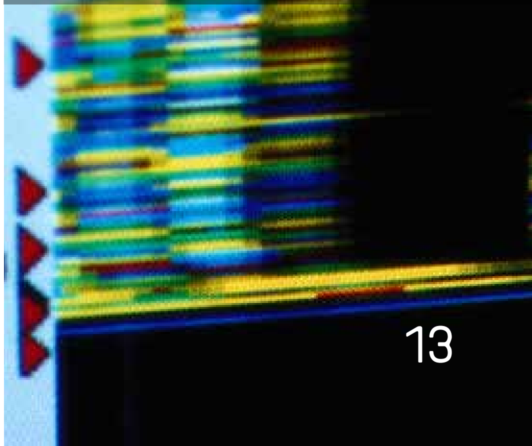
This Committee will focus on examining classical ethics ideologies (utilitarianism, etc) in light of AI and autonomous technologies.

Mixed Reality Committee

Mixed reality could alter our very notions of identity and reality over the next generation, as these technologies infiltrate more and more aspects of our lives, from work to education, from socializing to commerce. An AI backbone that would enable real-time personalization of this illusory world raises a host of ethical and philosophical questions, especially as the technology moves from headsets to much more subtle and integrated sensory enhancements. This Committee will work to discover the methodologies that could provide this future with an ethical skeleton and the assurance that the rights of the individual, including control over one's increasingly multifaceted identity, will be reflected in the encoding of this evolving environment.

> ***"Classical ethics methodologies have, to some degree, informed Artificial Intelligence and Autonomous Systems research since 1948, originating with Norbert Weiner's Cybernetics, the first values-driven methodology that sought to systematically study aspects of inherently biased values in artificial machine intelligence. By exploring ethics from several culturally diverse traditions and applying the thousands-year-old tradition of classical ethics to values-driven methodologies in ICTs a nd AI design we can achieve the goal of increasing human wellbeing for a positive future."***
>
> **Jared Bielby,**
> *(Chair, Classical Ethics Committee), Co-chair,*
> *International Center for Information Ethics*

> ***"Mixed Reality media combined with intelligent and autonomous technologies have the potential to rush us into a software-meditated world in which we see, hear and experience only what we want to see, hear and experience. This is why it is critical that technologists are trained in ethics so they can build and design technology that promotes and inspires our collective empathy, our work, ourselves and our society as a whole."***
>
> **Monique Morrow**
> *(Co-Chair, Mixed Reality Committee)*
> *CTO New Frontiers Engineering at Cisco*



NEW COMMITTEES[2]

Affective Computing

This Committee addresses the impact on individuals and society that autonomous systems capable of sensing, modeling, or exhibiting affective behavior such as emotions, moods, attitudes, and personality can produce. Affective computational and robotic artifacts have or are currently being developed for use in areas as diverse as companions, health, rehabilitation, elder and childcare, training and fitness, entertainment, and even intimacy. The ethical concerns surrounding human attachment and the overall impact on the social fabric may be profound and it is crucial that we understand the trajectories that affective autonomous systems may lead us on to best provide solutions that increase human wellbeing in line with innovation.

Policy: Effective Policymaking for Innovative Communities involving Artificial Intelligence and Autonomous Systems (EPICAIAS)

This Committee will:

1) explore how effective policymaking employing autonomous and intelligent technologies can be done in a rapidly changing world,
2) generate recommendations on what initiatives the private and public sector should pursue to positively impact individuals and society, and
3) illuminate newer models of policymaking both extant and experiment to support the innovation of AI/AS for shared human benefit.

"We need to provide ethical guidance regarding the appropriate design and use of affective computing within AI/AS to ensure that it does not violate the rights of users and society as a whole while at the same time assuring benefits to those who knowingly employ it for their own enjoyment and well being."

Ronald C. Arkin,
(Affective Computing Committee Co-Chair) Regents' Professor & Director of the Mobile Robot Laboratory; Associate Dean for Research & Space Planning, College of Computing Georgia Institute of Technology

"We need to be clear that the decision to use affect in intelligent systems has significant ethical ramifications. While there is clear utility to emotional systems in natural intelligence, both for individual control and social coordination, the confounding of emotion, suffering, and moral status in familiar natural examples of intelligence makes transparency concerning the role and nature of affect in AI particularly difficult and important."

Joanna Bryson
(Affective Computing Committee Co-Chair) Visiting Research Collaborator, Center for Information Technology Policy, Princeton University; Associate Professor, University of Bath, Intelligent Systems Research Group, Department of Computer Science

NEW COMMITTEES 3

ADDITIONAL QUOTES

"As mixed reality and associated technologies evolve, they will inevitably intermingle and become increasingly driven by AI. Just as "artificial intelligence" will eventually be seen simply as "intelligence," the convergence of these technologies that control perception and simulate reality will eventually be seen simply as "reality." That scenario raises a host of ethical concerns as well as questions about our very notions of self, identity, and reality."

Jay Iorio,
(Co-Chair, Mixed Reality Committee) Director of Innovation, IEEE Standards Association

---

"We need both policies and places in public service where we can collaborate with citizens and private sector partners on new ways of doing the business of public service better -- to include artificial intelligence and autonomous systems. The benefits of artificial intelligence to individual nations and the world are in the civilian domain, more so than any other domain."

Dr. David A. Bray,
(EPICAIAS Committee Co-Chair) Senior Executive & CIO for the FCC; Eisenhower Fellow to Taiwan and Australia; Harvard Visiting Executive In-Residence

---

"AI and autonomous systems will influence many aspects of life, business, health, and education. As a result, the ethical considerations are fraught with complexity. Standing as a bridge between policymakers and the commercial sector, The Global Initiative offers a uniquely impartial and balanced perspective in this global conversation. This work unites the public and private sectors, to the benefit of all"

Michael Krigsman,
Industry analyst and host of CXOTALK

Creative Direction and Design by Rich Silivanch

All painted work on pages 2, 5, 11, 14, 15 and 16 are copyright © 2016 Edward Povey & Tolar Schultz. All Rights Reserved. poveyandschultz.com