Version 1 - For Public Discussion

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

13 December 2016

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 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.

Table of Contents

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.

Executive Summary

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 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.

Executive Summary

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;

Executive Summary

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.

Executive Summary

- 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

Executive Summary

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.

Executive Summary

- 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.

Executive Summary

- 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*.

Executive Summary

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.

Executive Summary

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

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

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

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.

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 (AI/AS) regardless of whether they are physical robots (such as care robots or driverless cars) or software AIs (such as medical diagnosis systems, intelligent personal assistants, or algorithmic chat bots).

We are motivated by a desire to create ethical principles for 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 our intention that by identifying issues and draft recommendations these principles will eventually serve to underpin and scaffold future norms and standards within a new framework of ethical governance.

We have identified principles created by our Committee as well as additional principles reflected in the other Committees of The IEEE Global Initiative. We have purposefully structured our Committee and this document in this way to provide readers with a broad sense of the themes and ideals reflecting the nature of ethical alignment for these technologies as an introduction to our overall mission and work.

The following provides high-level guiding principles for potential solutions-by-design whereas other Committee sections address more granular issues regarding specific contextual, cultural, and pragmatic questions of their implementation.

1 General Principles

Principle 1 – Human Benefit

Issue:

How can we ensure that AI/AS do not infringe human rights?

Background

Documents such as The Universal Declaration of Human Rights,[i] the International Covenant for Civil and Political Rights,[ii] the Convention on the Rights of the Child,[iii] Convention on the Elimination of all forms of Discrimination against Women,[iv] Convention on the Rights of Persons with Disabilities[v] and the Geneva Conventions[vi] need to be fully taken into consideration by individuals, companies, research institutions, and governments alike to reflect the following concerns:

1. AI/AS should be designed and operated in a way that respects human rights, freedoms, human dignity, and cultural diversity.
2. AI/AS must be verifiably safe and secure throughout their operational lifetime.
3. If an AI/AS causes harm it must always be possible to discover the root cause (*traceability*) for said harm (*see also Principle 3 – Transparency*).

Candidate Recommendations

To best honor human rights, society must assure the safety and security of AI/AS to ensure they are designed and operated in a way that benefits humans:

1. Governance frameworks, including standards and regulatory bodies, should be established to oversee processes of assurance and of accident investigation to contribute to the building of public trust in AI/AS.
2. A methodology is also needed for translating existing and forthcoming legal obligations into informed policy and technical considerations.

Further Resources

The following documents/organizations are provided both as references and examples of the types of work that can be emulated, adapted, and proliferated, regarding ethical best practices around AI/AS to best honor human rights:

- The Universal Declaration of Human Rights.
- The International Covenant on Civil and Political Rights, 1966.
- The International Covenant on Economic, Social and Cultural Rights, 1966.
- The International Convention on the Elimination of All Forms of Racial Discrimination, 1965.

1 General Principles

- The Convention on the Rights of the Child.
- The Convention on the Elimination of All Forms of Discrimination against Women, 1979.
- The Convention on the Rights of Persons with Disabilities, 2006.
- The Geneva Conventions and additional protocols, 1949.
- IRTF's Research into Human Rights Protocol Considerations.
- The UN Guiding Principles on Business and Human Rights, 2011.

Principle 2 – Responsibility

Issue:

How can we assure that AI/AS are accountable?

Background

The programming and output of AI/AS are often not discernible by the general public. Based on the cultural context, application, and use of AI/AS, people and institutions need clarity around the manufacture of these systems to avoid potential harm. Additionally, manufacturers of these systems must be able to provide programmatic-level accountability proving why a system operates in certain ways to address legal issues of culpability, and to avoid confusion or fear within the general public.

Candidate Recommendations

To best address issues of responsibility:

1. Legislatures/courts should clarify issues of responsibility, culpability, liability, and accountability for autonomous and intelligent systems where possible during development and deployment (to free manufacturers and users to understand what their rights and obligations should be).
2. Designers and developers of autonomous and intelligent systems should remain aware of, and take into account when relevant, the diversity of existing cultural norms among the groups of users of these AI/AS.
3. Multi-stakeholder ecosystems should be developed to help create norms where they don't exist because AI/AS-oriented technology and their impacts are too new (including representatives of civil society, law enforcement, insurers, manufacturers, engineers, lawyers, etc.).
4. Systems for registration should be created by producers/users of autonomous systems (capturing key, high-level parameters), including:
    - Intended use
    - Training data (if applicable)
    - Sensors/real world data sources
    - Algorithms
    - Process graphs
    - Model features (at various levels)
    - User interfaces
    - Actuators/outputs
    - Optimization goal/loss function/ reward function

Further Resources

- (In relation to Candidate Recommendation #3) Sciencewise – The U.K. national center for public dialogue in policymaking involving science and technology issues.

Principle 3 – Transparency

Issue:

How can we ensure that AI/AS are transparent?

Background

A key concern over autonomous systems is that their operation must be transparent to a wide range of stakeholders for different reasons (noting that the level of transparency will necessarily be different for each stakeholder). Stated simply, a *transparent* AI/AS is one in which it is possible to discover how and why the system made a particular decision, or in the case of a robot, acted the way it did.

AI/AS will be performing tasks that are far more complex and impactful than prior generations of technology, particularly with systems that interact with the physical world, thus raising the potential level of harm that such a system could cause. Consider AI/AS that have real consequences to human safety or wellbeing, such as medical diagnosis AI systems, or driverless car autopilots; systems such as these are *safety-critical* systems.

At the same time, the complexity of AI/AS technology itself will make it difficult for users of those systems to understand the capabilities and limitations of the AI systems that they use, or with which they interact, and this opacity, combined with the often-decentralized manner in which it is developed, will complicate efforts to determine and allocate responsibility when something goes wrong with an AI system. Thus, lack of transparency both increases the risk and magnitude of harm (users not understanding the systems they are using) and also increases the difficulty of ensuring accountability.

Transparency is important to each stakeholder group for the following reasons:

1. For users, transparency is important because it builds trust in the system, by providing a simple way for the user to understand what the system is doing and why.
2. For validation and certification of an AI/AS, transparency is important because it exposes the system's processes for scrutiny.
3. If accidents occur, the AS will need to be transparent to an accident investigator, so the internal process that led to the accident can be understood.
4. Following an accident, judges, juries, lawyers, and expert witnesses involved in the trial process require transparency to inform evidence and decision-making.
5. For disruptive technologies, such as driverless cars, a certain level of transparency to wider society is needed in order to build public confidence in the technology.

1 General Principles

Candidate Recommendation

Develop new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined. For designers, such standards will provide a guide for self-assessing transparency during development and suggest mechanisms for improving transparency. (The mechanisms by which transparency is provided will vary significantly, for instance (1) for users of care or domestic robots a why-did-you-do-that button which, when pressed, causes the robot to explain the action it just took, (2) for validation or certification agencies the algorithms underlying the AI/AS and how they have been verified, (3) for accident investigators, secure storage of sensor and internal state data, comparable to a flight data recorder or black box.)

Further Resources

- Transparency in Safety-Critical Systems, Machine Intelligence Research Institute, August 2013.
- M Scherer, Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies, May 2015.
- See section on Decision Making Transparency in the Report of the U.K. House of Commons Science and Technology Committee on Robotics and Artificial Intelligence, 13 September 2016.

Principle 4 – Education and Awareness

Issue:

How can we extend the benefits and minimize the risks of AI/AS technology being misused?

Background

In an age where these powerful tools are easily available, there is a need for new kind of education for citizens to be sensitized to risks associated with the misuse of AI/AS. Such risks might include hacking, "gaming," or exploitation (e.g., of vulnerable users by unscrupulous manufacturers).

Candidate Recommendations

Raise public awareness around the issues of potential AI/AS misuse in an informed and measured way by:

1. Providing ethics education and security awareness that sensitizes society to the potential risks of misuse of AI/AS.
2. Delivering this education in new ways, beginning with those having the greatest impact that also minimize generalized (e.g., non-productive) fear about AI/AS (e.g., via accessible science communication on social media such as Facebook or YouTube).
3. Educating law enforcement surrounding these issues so citizens work collaboratively with them to avoid fear or confusion (e.g., in the same way police officers have given public safety lectures in schools for years, in the near future they could provide workshops on safe AI/AS).

Further Resources

- (In relation to Candidate Recommendation #2) Wilkinson, Clare, and Emma Weitkamp. *Creative Research Communication: Theory and Practice*. Manchester University Press, 2016.

2 Embedding Values Into Autonomous Intelligent Systems

Society does not have universal standards or guidelines to help embed human norms or moral values into autonomous intelligent systems (AIS) today. But as these systems grow to have increasing autonomy to make decisions and manipulate their environment, it is essential they be designed to adopt, learn, and follow the norms and values of the community they serve, and to communicate and explain their actions in as transparent and trustworthy manner possible, given the scenarios in which they function and the humans who use them.

The conceptual complexities surrounding what "values" are make it currently difficult to envision AIS that have computational structures directly corresponding to values. However, it is a realistic goal to embed explicit norms into such systems, because norms can be considered instructions to act in defined ways in defined contexts. A community's network of norms as a whole is likely to reflect the community's values, and AIS equipped with such a network would therefore also reflect the community's values, even if there are no directly identifiable computational structures that correspond to values.

To address this need, our Committee has broken the broader objective of embedding values into these systems into three major goals:

1. Identifying the norms and eliciting the values of a specific community affected by AIS.
2. Implementing the norms and values of that community within AIS.
3. Evaluating the alignment and compatibility of those norms and values between the humans and AIS within that community.

2 Embedding Values Into Autonomous Intelligent Systems

Pursuing these three goals represents an iterative process that is contextually sensitive to the requirements of AIS, their purpose, and their users within a specific community. It is understood that there will be clashes of values and norms when identifying, implementing, and evaluating these systems (a state often referred to as "moral overload"). This is why we advocate for a stakeholder-inclusive approach where systems are designed to provide transparent signals (such as explanations or inspection capabilities) about the specific nature of their behavior to the various actors within the community they serve. While this practice cannot always eliminate the possible data bias present in many machine-learning algorithms, it is our hope that the proactive inclusion of users and their interaction with AIS will increase trust in and overall reliability of these systems.

Identifying Norms and Values for Autonomous Intelligent Systems

Issue:

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

Background

If machines enter human communities as autonomous agents, then those agents will be expected to follow the community's social and moral norms. A necessary step in enabling machines to do so is to identify these norms. Whereas laws are formalized and therefore relatively easy to identify, social and moral norms are more difficult to ascertain, as they are expressed through behavior, language, customs, cultural symbols, and artifacts. Moreover, communities (from families to whole nations) differ to various degrees in the norms they follow. So embedding norms in AIS requires a clear delineation of the community in which AIS are to be deployed. Further, even within the same community, different types of AIS will demand different sets of norms. The relevant norms for self-driving vehicles, for example, will differ greatly from those for robots used in healthcare.

Candidate Recommendation

We acknowledge that generating a universal set of norms/values that is applicable for all autonomous systems is not realistic. Instead, we recommend to first identify the sets of norms that AIS need to follow in specific communities and for specific tasks. Empirical research involving multiple disciplines and multiple methods should investigate and document these numerous sets of norms and make them available for designers to implement in AIS.

Further Resources

This book describes some of the challenges of having a one-size-fits-all approach to embedding human values in autonomous systems: Wallach, Wendell and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2008.

2 Embedding Values Into Autonomous Intelligent Systems

Issue:

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

Background

An autonomous system is often built with many constraints and goals in mind. These include legal requirements, monetary interests, and also social and moral values. Which constraints should designers prioritize? If they decide to prioritize social and moral norms of end users (and other stakeholders), how would they do that?

Candidate Recommendation

Our recommended best practice is to prioritize the values that reflect the shared set of values of the larger stakeholder groups. For example, a self-driving vehicle's prioritization of one factor over another in its decision making will need to reflect the priority order of values of its target user population, even if this order is in conflict with that of an individual designer, manufacturer, or client. For example, the Common Good Principle[vii] could be used as a guideline to resolve differences in the priority order of different stakeholder groups.

We also recommend that the priority order of values considered at the design stage of autonomous systems have a clear and explicit rationale. Having an explicitly stated rationale for value decisions, especially when these values are in conflict with one another, not only encourages the designers to reflect on the values being implemented in the system, but also provides a grounding and a point of reference for a third party to understand the thought process of the designer(s). The Common Good Principle mentioned above can help formulate such rationale.

We also acknowledge that, depending on the autonomous system in question, the priority order of values can dynamically change from one context of use to the next, or even within the same system over time. Approaches such as interactive machine learning (IML), or direct questioning and modeling of user responses can be employed to incorporate user input into the system. These techniques could be used to capture changing user values.

Further Resources

- Markkula Center for Applied Ethics, The Common Good. Idea of the common good decision-making was introduced here.
- Van den Hoven, Jeroen, Engineering and the Problem of Moral Overload. *Science and Engineering Ethics* 18, no. 1 (March 2012): 143-155.
- One of the places where differences in human moral decision-making and changes in priority order of values for autonomous systems are documented is a series of poll results published by the Open Roboethics initiative. In particular, see these poll results on care robots.

2 Embedding Values Into Autonomous Intelligent Systems

Issue:

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

Background

Autonomous intelligent systems, compared to traditional systems, are sometimes discussed as a new type of species—called the new ontological category,[vii] according to literature in human-robot interaction—because of the manner in which humans perceive, interact with, and psychologically respond to them. For example, numerous studies have documented the way in which humans willingly follow even the strangest of requests from a robot, demonstrating the impact these systems can have on our decision-making and behavior (see for example, Robinette, Paul, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner, "Overtrust of Robots in Emergency Evacuation Scenarios,"[viii] 2016 ACM/IEEE International Conference on Human-Robot Interaction). Hence, it is important to be aware of possible use of the systems for the purposes of manipulation.

In addition, various aspects of these systems can be designed to instill bias into other users, whether intended or not. The sources of bias can span from the way a system senses the world (e.g., can the system detect a person missing an arm or does it assume all humans have two arms?), to how it processes and responds to the sensed information (e.g., does the system respond to people of different ethnicity, gender, race, differently?), as well as what it looks like. Details of an interactive autonomous system's behavior can have far-reaching consequences, such as reinforcement of gender, ethnic, and other biases (see for example, Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai, "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings,"[ix] Cornell University Library, arXiv:1607.06520, July 21, 2016.)

Moreover, while deciding which values and norms to prioritize, we call for special attention to the interests of vulnerable and under-represented populations, such that these user groups are not exploited or disadvantaged by (possibly unintended) unethical design. While traditionally the term *vulnerable populations* refers to disadvantaged sub-groups within human communities—including but not limited to children, older adults, prisoners, ethnic minorities, economically disadvantaged, and people with physical or intellectual disabilities—here we also include populations who may not be traditionally considered a member of vulnerable populations, but may be so in the context of autonomous intelligent systems. For example, riders in autonomous vehicles, or factory workers using a 400-pound high-torque robot, who would not otherwise be vulnerable under the traditional definition, become vulnerable in the use contexts due to the user's reliance on the system or physical disadvantage compared to the high-powered machinery.

2 Embedding Values Into Autonomous Intelligent Systems

Candidate Recommendation

It is important to acknowledge that it is easy to have built-in biases in autonomous systems. For example, a system that depends on face recognition trained entirely on Caucasian faces may work incorrectly or not at all on people with non-Caucasian skin tones or facial structures. This renders the system to be perceived as discriminatory, whether it was designed with such intent or not. These biases can also stem from the values held by the designer. We can reduce the incidence of such unintended biases by being more aware of the potential sources of these biases. We posit that being aware of this particular issue and adopting more inclusive design principles can help with this process. For example, systems that can sense persons of different races, ethnicities, genders, ages, body shapes, or people who use wheelchairs or prosthetics, etc.

We also highlight that this concern delves into the domain of ongoing research in human-robot interaction and human-machine interaction. To what extent and how do built-in biases change the course of robot interaction with human users? What dynamic and longitudinal effect do they have on the users and the society? How does a robot's morphology in different use cases affect target user groups? These are all open research questions for which we do not yet have clear answers. Since there is no clear understanding of the nature of these biases and their alignment with human values, we recommend conducting research and educational efforts to resolve these open questions and to address these issues in a participatory way by introducing into the design process members of the groups who may be disadvantaged by the system.

In particular, vulnerable populations are often one of the first users of autonomous systems. In designing for these populations, we recommend designers familiarize themselves with relevant resources specific to the target population. We also note that a system can have multiple end users, each of which may demand a conflicting set of values. We recommend designers be aware of such conflicts and be transparent in addressing these conflicting value priorities as suggested in the above-mentioned issue. AIS are usually subject to a multiplicity of norms and values that may conflict with each other.

Therefore, we strongly encourage the inclusion of intended stakeholders in the entire engineering process, from design and implementation to testing and marketing, as advocated for example in disability studies literature (see "Nothing About Us Without Us" in the Further Resources below).

A number of institutions have established connections with communities of a particular vulnerable population (e.g., University of Washington's DO-IT program). However, there is no one voice that represents all vulnerable populations. Hence, we recommend designers and practitioners reach out to communities of interest and relevant advocacy groups.

We also recommend, especially when designing for dynamically vulnerable populations, that designers take on an interdisciplinary approach and involve relevant experts or advisory group(s) into the design process. Thus, designers of AIS should work together with behavioral scientists

2 Embedding Values Into Autonomous Intelligent Systems

and members of the target populations to systematically study population norms, expectations, concerns, and vulnerabilities. We also encourage designers to include regulators and policymakers in this process as well, noting that shaping regulation and policy is an integral part of guiding the development and deployment of autonomous systems in a desirable direction.

Further Resources

- Asaro, P. "Will BlackLivesMatter to RoboCop?" We Robot, 2016.
- Riek, L. D. and D. Howard. A Code of Ethics for the Human-Robot Interaction Profession. We Robot, 2014.
- Winfield, A. Robots Should Not Be Gendered (blog), 2016.
- Whitby, Blay. "Sometimes It's Hard to Be a Robot: A Call for Action on the Ethics of Abusing Artificial Agents." *Interacting with Computers* 20, no. 3 (2008): 326-333.
- Federal Trade Commission. Privacy Online: Fair Information Practices in the Electronic Marketplace: A Federal Trade Commission Report to Congress. 2000.
- Riek, Laurel D. "Robotics Technology in Mental Health Care." *Artificial Intelligence in Behavioral Health and Mental Health Care*, (2015): 185-203.
- Charlton, James I. *Nothing About Us Without Us: Disability Oppression and Empowerment*, University of California Press, 2000.
- Shivayogi, P. "Vulnerable Population and Methods for Their Safeguard." *Perspectives in Clinical Research*, January-March (2013): 53-57.

Embedding Norms and Values in Autonomous Intelligent Systems

Issue:

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.

Background

The prospect of developing computer systems that are sensitive to human norms and values and factoring these issues into making decisions in morally or legally significant situations has intrigued science fiction writers, philosophers, and computer scientists alike. Modest efforts to realize this worthy goal in limited or bounded contexts are already underway. This emerging field of research goes under many names including: machine morality, machine ethics, moral machines, value alignment, computational ethics, artificial morality, safe AI, and friendly AI. Basic notions can be found in books such as Allen, C., and W. Wallach. *Moral Machines.*[x] *Teaching Robots Right from Wrong*. Oxford University Press, 2010.

Computers and robots already instantiate values in their choices and actions, but these values are programmed or designed by the engineers that build the systems. Increasingly, autonomous systems will encounter situations that their designers cannot anticipate, and will require algorithmic procedures to select the better of two or more possible courses of action. Some of the existing experimental approaches to building moral machines are top-down. In this sense the norms, rules, principles, or procedures are used by the system to evaluate the acceptability of differing courses of action or as moral standards or goals to be realized.

Recent breakthroughs in machine learning and perception will enable researchers to explore bottom-up approaches—in which the AI system learns about its context and about human values—similar to the manner in which a child slowly learns which forms of behavior are safe and acceptable. Of course a child can feel pain and pleasure, empathize with others, and has other capabilities that AI system cannot presently imitate. Nevertheless, as research on autonomous systems progresses, engineers will explore new ways to either simulate these capabilities, or build alternative mechanisms that fulfill similar functions.

2 Embedding Values Into Autonomous Intelligent Systems

Candidate Recommendation

Research on this front should be encouraged. Advances in data collection, sensor technology, pattern recognition, machine learning, and integrating different kinds of data sets will enable creative, new approaches for ensuring that the actions of AI systems are aligned with the values of the community in which they operate. Progress toward building moral machines may well determine the safety and trustworthiness of increasingly autonomous AI systems.

Further Resources

- Allen, C., and W. Wallach. *Moral Machines: Teaching Robots Right from Wrong.* Oxford University Press, 2010.
- Anderson, M., and S. Anderson (eds.). *Machine Ethics.* Cambridge University Press, 2011.
- Abney, K., G. Bekey, and P. Patrick. *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT Press, 2011.
- RC Arkin, P Ulam, AR Wagner, Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception, Proceedings of the IEEE 100 (3), 571-589

Evaluating the Alignment of Norms and Values between Humans and AIS

Issue:

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

Background

If a community's systems of norms (and their underlying values) has been identified, and if this process has successfully guided the implementation of norms in AIS, then the third step in value embedding must take place: rigorous testing and evaluation of the resulting human-machine interactions regarding these norms.

An intuitive criterion in these evaluations might be that the norms embedded in AIS should correspond closely to the human norms identified in the community—that is, AIS should be disposed to behave the same way that people expect each other to behave. However, for a given community and a given AIS task and use context, AIS and humans may not have identical, but rather compatible, sets of norms. People will have some unique expectations for humans that they don't have for machines (e.g., norms governing the expression of emotions, as long as machines don't have, or clearly express, emotions), and people will have some unique expectations of AIS that they don't have for humans (e.g., that the machine will destroy itself if it can thereby prevent harm to a human). The norm identification process must document these structural relations (similarities as well as differences) between human and AIS norms, and in evaluating these relations, the goal of *compatibility* may be preferred over that of *alignment*, which suggests primarily a similarity structure.

In addition, more concrete criteria must be developed that indicate the quality of human-machine interactions, such as human approval and appreciation of AIS, trust in AIS, adaptability of AIS to humans users, and human benefits in the presence or influence of AIS. Evaluation of these and other criteria must occur both before broad deployment and throughout the life cycle of the system. Assessment before deployment would best take place in systematic test beds that allow human users (from the defined community) to engage safely with AIS (in the defined tasks) and enable assessment of approval, trust, and related variables. Examples include the Tokku testing zones in Japan.[xi]

2 Embedding Values Into Autonomous Intelligent Systems

Candidate Recommendation

The success of implementing norms in AIS must be rigorously evaluated by empirical means, both before and throughout deployment. Criteria of such evaluation will include compatibility of machine norms and human norms (so-called *value alignment* or *compliance*, depending on the nature of the norms), human approval of AIS, and trust in AIS, among others. Multiple disciplines and methods should contribute to developing and conducting such evaluation, such as extensive tests (including adversarial ones), explanation capabilities to reconstruct AIS inner functioning, natural language dialog between AIS and humans (including deep question answering), and context awareness and memory (to handle repeated evaluations).

Issue:

Achieving a correct level of trust between humans and AIS.

Background

Development of autonomous systems that are worthy of our trust is challenged due to the current lack of transparency and verifiability regarding these systems for users. For this issue, we explore two levels at which transparency and verifiability are useful and often necessary. A first level of transparency relates to the information conveyed to the user while an autonomous system interacts with the user. A second level has to do with the possibility to evaluate the system as a whole by a third party (e.g., regulators, society at large, and post-accident investigators).

In the first level, consider for example the case of robots built to interact with people. The robots should be designed to be able to communicate what they are about to perform and why as the actions unfold. This is important in establishing an appropriate level of trust with the user. While a system that a user does not trust may never be used, a system that is overly trusted can negatively affect the user as well based on the perception of the particular system or similar types of systems by the society. Unlike humans who naturally use verbal and nonverbal behaviors to convey trust-based information to those around them, the mode and the content of communicative behaviors toward or from an autonomous system are features that would be absent if not for the explicit implementation by the designers. Designing systems that are worthy of our trust necessarily includes making these explicit design decisions. As with people, trust is built over time, through repeated interactions, so AIS must be equipped with context awareness and memory capabilities.

Candidate Recommendation

Transparency and verifiability are necessary for building trust in AIS. We recommend that AIS come equipped with a module assuring some level of transparency and verifiability. Technological solutions to address the issue of transparency and instilling the right level of trust in the users is an open area of research. Trust

2 Embedding Values Into Autonomous Intelligent Systems

is also a dynamic variable in human-machine interaction; the level of trust a user may have with a system tends to change over time. Coupled with the dynamic nature of trust in autonomous systems is our known tendency to overly trust technology beyond its capabilities. With systems that have been commercialized, for example, users often assume a minimum level of reliability and trustworthiness of the system from the onset.

Hence, even when a system is delivered with a written disclaimer outlining its conditions of use, it is often naïve to assume that the disclaimer alone can protect the interests of both the manufacturer/developer and users. In addition to communicating the limitations and capabilities of the system to the users, we recommend autonomous systems to be designed with features that prevent users from operating the system outside a known, safe, and appropriate range of conditions of use, including conditions that depend on user behavior. We also recommend evaluation of the system's design with the user's perception of their role in mind (e.g., operator versus user of the system), such that the system's interaction with the user is in alignment with the role that is expected of the user.

In addition, one can design communicative and behavioral features of a system to serve as interactive real-time disclaimers, such that the user is informed of significant changes to the system's level of confidence on a proposed solution for the task to be performed, which can change from one moment or situation to the next. Systems that lack such features can result in not only ineffective interaction with the user—introducing a point of miscommunication, for example—but also risk the safety and wellbeing of the user and others. This also makes it more challenging for a user to diagnose the reasons why a system may be behaving in a certain way, and to detect when malfunctions occur.

Issue:

Third-party evaluation of AIS's value alignment.

Background

The second level of transparency, as stated above, is needed to evaluate a system as a whole by a third party (e.g., regulators, society at large, and post-accident investigators).

In this second category, there are concerns regarding the increasing number of autonomous systems that rely on, or include, AI/machine-learning techniques inherently lacking transparency and verifiability. Discussions on this topic include: the nature and possible bias of the data sets used to train a machine-learning system that is often not accessible by the public, details of the algorithm used to create the final product, the specifications on the final product's efficacy and performance, and the need to consider the scenario where AIS will be used when evaluating their adherence to relevant human values. While acknowledging the usefulness and potential for

2 Embedding Values Into Autonomous Intelligent Systems

these systems, it is a serious concern that even the designers and programmers involved cannot verify or guarantee reliability, efficacy, and value alignment of the final system. A further problem is that there is no agreed-upon method, process, or standards for validating and certifying the adherence of AIS to desired human norms and values.

Candidate Recommendation

With regards to our concern on the transparency between a system as a whole and its evaluator (e.g., regulator), we recommend that designers and developers alike document changes to the systems in their daily practice. A system with the highest level of traceability would contain a black-box-like module such as those used in the airline industry, that logs and helps diagnose all changes and behaviors of the system. Such practice, while it does not fully address the need for transparency of a number of popular machine-learning approaches, allows one to trace back to the sources of problems that may occur and provide a mechanism with which a faulty behavior of a system can be diagnosed.

As more human decision-making is delegated to autonomous systems, we expect there to be an increasing need for rationale and explanation as to how the decision was reached by the algorithm. In this respect, a relevant regulation is the European Union's new General Data Protection Regulation (GDPR)[xiv], adopted on April 2016 and scheduled to take effect in 2018. The GDPR states that, in regards to automated decisions based on personal data, individuals have a right to "an explanation of the [algorithmic] decision reached after such assessment and to challenge the decision." While the development of an algorithm that is able to explain its behavior is an open research topic, there are algorithms that are more transparent than others, such as logic-based AI that provide more transparency than machine-learning AI, and more coherence between the output behavior of a system and its inner functioning. Winfield, Blum, and Liu's work on consequence engine[xv], for example, utilizes a simulator to predict and evaluate the consequences of an artificial agent's possible next actions in order to decide the right course of action, making the agent's decision-making process easy to examine and validate. In the absence of an adequate alternative, it is imperative that designers be aware of the need for transparency and strive to increase it in the algorithms they design and implement into autonomous systems.

We also recommend that regulators define, together with users, developers, and designers, a minimum level of value alignment and compliance, and suitable capabilities for this to be checked by a third party, in order for AIS to be deployed.

Finally, we recommend to define criteria to define AIS as trustworthy. These criteria will depend on a machine's expected tasks and context of use, as well as the users' vulnerabilities (we expect that more-vulnerable-user categories will require more stringent criteria).

2 Embedding Values Into Autonomous Intelligent Systems

Further Resources

- Goodman, B., and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'," Cornell University Library, arXiv: 1606.08813, August 31, 2016.
- Winfield, A. F. T., C. Blum, and W. Liu, "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection," *Advances in Autonomous Robotics Systems*, Lecture Notes in Computer Science Volume 8717, (2014): 85-96. Eds. Mistry M, Leonardis A, Witkowski M and Melhuish C, Springer, 2014.

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 rights, as defined in the Universal Declaration of Human Rights, as a primary form of human values. Therefore, we strongly believe that values-aligned design methodology should become an essential focus for the modern AI/AS organization.

Values-aligned system design puts human flourishing at the center of IT development efforts. It recognizes that machines should serve humans and not the other way around. It aims to create sustainable systems that are thoroughly scrutinized for social costs and advantages that will also increase economic value for organizations by embedding human values in design.

To help achieve these goals, technologists will need to embrace transparency regarding their products to increase end user trust. The proliferation of values-based design will also require a change of current system development approaches for organizations, including a commitment to the idea that innovation should be defined by human-centricity versus speed to market.

The process of utilizing multiple ethical approaches to provably aligned end user values will provide a key competitive differentiator in the algorithmic economy by prioritizing respect for individuals above exponential growth. Progressive organizations honoring values-based design will lead the creation of standards and policies that inform end users and other stakeholders, providing conscious consent for the use of their intelligent and autonomous technology.

Section 1 – Interdisciplinary Education and Research

Integrating applied ethics into education and research to address the issues of Artificial Intelligence and Autonomous Systems (AI/AS) requires an interdisciplinary approach, bringing together humanities, social sciences, science and engineering disciplines.

Issue:

Ethics is not part of degree programs.

Background

AI engineers and design teams too often fail to discern the ethical decisions that are implicit in technical work and design choices, or alternatively, treat ethical decision-making as just another form of technical problem solving. Moreover, technologists often struggle with the imprecision and ambiguity inherent in ethical language, which cannot be readily articulated and translated into the formal languages of mathematics, and computer programming associated with algorithms and machine learning. Thus, ethical issues can easily be rendered invisible or inappropriately reduced/simplified in the context of technical practice. This originates in the fact that Engineering programs do not often require coursework, training, or practical experience in applied ethics. A methodology for bridging the need of a truly interdisciplinary and intercultural education of the intricacies of technology and its effects on human society for the engineers who develop said technologies is required especially in regard to the immediacy ethical considerations of AI/AS

Candidate Recommendation

Ethics and ethical reflection need to be a core subject for engineers and technologists beginning at University level and for all advanced degrees. By making students sensitive to ethically aligned design issues before they enter the workplace, they can implement these methodologies in a cross-disciplinary fashion in their jobs. It is also important that these courses not be contained solely within an ethics or philosophy department but infused throughout arts, humanities and technology programs. Human values transcend all academic areas of focus.

We also recommend establishing an intercultural and interdisciplinary curriculum that is informed by ethicists, scientists, philosophers, psychologists, engineers and subject matter experts from a variety of cultural backgrounds that can be used to inform and teach aspiring engineers (post-secondary) about the relevance

3 Methodologies to Guide Ethical Research and Design

and impact of their decisions in designing AI/AS technologies. Even more critical is the priority to introduce a methodology for bridging the need for a truly interdisciplinary and intercultural education of the intricacies of technology into primary and secondary education programs. These courses should be part of the technical training and engineering development methodologies so that ethics becomes naturally part of the design process.

Further Resources

- A good example of such cross pollination can be found in the work and workshops organized by Ben Zevenbergen and Corinne Cath of the Oxford Internet Institute. The following workshop outcomes paper addresses some ethical issues in engineering from a multi-disciplinary perspective: Philosophy Meets Internet Engineering: Ethics in Networked Systems Research.
- The White House report on 'Preparing for the Future of AI' makes several recommendations on how to ensure that AI practitioners are aware of ethical issues by providing them with ethical training.
- The French Commission on the Ethics of Research in Digital Sciences and Technologies (CERNA) recommends including ethics classes in doctoral degree.
- Companies should also be encouraged to mandate consideration of ethics at the pre-product design stage, as was done by Lucid AI.

Issue:

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

Background

Not enough models exist for bringing engineers and designers in contact with ethicists and social scientists, both in academia and industry, so that meaningful interdisciplinary collaboration can shape the future of technological innovation.

Candidate Recommendation

This issue, to a large degree, relates to funding models, which limit cross-pollination between disciplines (see below). To help bridge this gap, more networking and collaboration between ethicists and technologists needs to happen in order to do the "translation work" between the worlds of investigating the social implications of technology and its actual design. Even if reasoning methods and models may differ across disciplines, sharing actual experience and knowhow is central to familiarize technologists with ethical approaches in other disciplines (e.g., medicine, architecture). Global professional organizations should devote specific access to resources (websites, MOOCS etc.) for sharing experience and methodologies.

3 Methodologies to Guide Ethical Research and Design

Further Resources

- Value Sensitive Design as described by Batya Friedman as well as Value-based Design as proposed by Sarah Spiekermann, both foresee the integration of value analysis into system design. Values are identified by senior executives and innovation team members; potentially supported by a Chief Officer devoted to this task. Then the identified values are conceptually analyzed and broken down to identify ways of system integration. Both approaches can be studied in more detail in Sarah Spiekermann's book, Ethical IT Innovation: A Value-Based System Design Approach.
- The methodology developed by the Internet Research Task Force's Human Rights Protocol Research Group (HRPC) is another example of a relevant methodology. Their guidelines provide us with an example of how human values, ethical or otherwise, relate and can be translated to Internet technology. Their website details how these values can be used in technology (both in language and in process) to fit into the Internet Engineering Task Force/ Internet Research Task Force (IETF/IRTF) engineering processes. In short, relevant values are identified on the basis of the Universal Declaration of Human Rights. These different rights are broken down into their various components and then matched to technical concepts in the process of an Internet protocol design. By combining the different technical concepts as they match different human rights components - protocol designers can approximate human rights through their work.

Issue:

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

Background

A responsible approach to embedded values (both as bias and as value by design) in ICTs, algorithms and autonomous systems will need to differentiate between culturally distinctive values (i.e. how do different cultures view privacy, or do they at all? And how do these differing presumptions of privacy inform engineers and technologists and the technologies designed by them?). Without falling into ethical relativism, it is critical in our international IEEE Global Initiative to avoid only considering western influenced ethical foundations. Other cultural ethical/moral, religious, corporate and political traditions need to be addressed, as they also inform and bias ICTs and autonomous systems.

Candidate Recommendation

Establish a leading role for Intercultural Information Ethics[xvi] (IIE) practitioners in value-by-design ethics committees informing technologists, policy makers and engineers. Clearly demonstrate through examples how cultural bias informs not only information flows and information systems but also algorithmic decision-making and value by design.

3 Methodologies to Guide Ethical Research and Design

Further Resources

- The work of David, et al. (2006) and Bielby (2015) has been guiding in this field "Cultural values, attitudes, and behaviors prominently influence how a given group of people views, understands, processes, communicates, and manages data, information, and knowledge."

- Pauleen, David J., et al. "Cultural Bias in Information Systems Research and Practice: Are You Coming From the Same Place I Am?." *Communications of the Association for Information Systems* 17.1 (2006): 17.
- Bielby, Jared. "Comparative Philosophies in Intercultural Information Ethics," *Confluence: Online Journal of World Philosophies* vol. 2, no. 1, (2015): 233-253.

Section 2 – Business Practices and AI

Businesses are eager to develop and monetize AI/AS but there is little supportive structure in place for creating ethical systems and practices around its development or use.

Issue:

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

Background

There is a need to create value-based ethical culture and practices for the development and deployment of products based on Autonomous Systems.

Candidate Recommendation

The building blocks of such practices include top-down leadership, bottom-up empowerment, ownership and responsibility, and need to consider system deployment contexts and/or ecosystems. The institution of such cultures would accelerate the adoption of the other recommendations associated within this section focused on Business Practices.

Further Resources

- The website of the Benefit Corporations (B Corporations) provides a good overview of a range of companies that personify this type of culture.

Issue:

Lack of values-aware leadership.

Background

Technology leaders give innovation teams and engineers too little or no direction on what human values should be respected in the design of a system. The increased importance of AI/AS systems in all aspects of our wired societies further accelerates the needs for value-aware leadership in AI/AS development.

Candidate Recommendations

Chief Values Officers

Companies need to create roles for senior-level marketers, ethicists or lawyers who can pragmatically implement ethically aligned design. A precedent for this type of methodological adoption comes from Agile Marketing [xvii] whose origin began in open source and engineering circles. Once the business benefits of Agile were clearly demonstrated to senior management, marketers began to embrace these

3 Methodologies to Guide Ethical Research and Design

methodologies. In today's algorithmic economy, organizations will quickly recognize the core need to identify and build to end-user values. A precedent for this new type of leader can be found in the idea of a Chief Values Officer created by Kay Firth-Butterfield. [xviii]

However, ethical responsibility should not be delegated to chief values officers. CVOs can support the creation of ethical knowledge in companies, but in the end all members of an innovation team will need to act responsibly throughout the design process.

Embedded Industry-Wide CSR
Given the need for engineers to understand intimately the cultural context and ethical considerations of design decisions, particularly as technologies afford greater levels of power, autonomy and surveillance, corporations should make a deliberate effort to ground engineering practice in authentic cultural inquiry. By creating the exemplar guidelines to enable every corporation to set up community-centered CSR efforts, companies can dedicate specific engineering resources to local problems using technology innovation for social good.

Further Resources

- As an example to emulate for embedded industry-wide Corporate Social Responsibility CSR, we recommend the Gamechangers 500 Index.

Issue:

Lack of empowerment to raise ethical concerns.

Background

Engineers and design teams are neither socialized nor empowered to raise ethical concerns regarding their designs, or design specifications, within their organizations. Considering the widespread use of AI/AS and the unique ethical questions it raises, these need to be identified and addressed from their inception.

Candidate Recommendation

Code of Conduct
In a paradigm that more fully recognizes and builds to human values, employees should be empowered to raise concerns around these issues in day to day professional practice, not just in extreme emergency circumstances such as whistleblowing. New organizational processes need to be implemented within organizations that broaden the scope around professional ethics and design as AI/AS has raised issues that do not fit the existing paradigms. New categories of considerations around these issues need to be accommodated as AI/AS have accelerated the need for new forms of Code of Conducts, so individuals feel proactively empowered to share their insights and concerns in an atmosphere of trust.

3 Methodologies to Guide Ethical Research and Design

Example: The British Computer Society (BCS) [xix] code of conduct holds that individuals have to: "a) have due regard for public health, privacy, security and wellbeing of others and the environment. b) have due regard for the legitimate rights of Third Parties*. c) conduct your professional activities without discrimination on the grounds of sex, sexual orientation, marital status, nationality, color, race, ethnic origin, religion, age or disability, or of any other condition or requirement. d) promote equal access to the benefits of IT and seek to promote the inclusion of all sectors in society wherever opportunities arise."

Further Resources

- The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights mitigates the issue surrounding the lack of empowerment to raise ethical concerns by suggesting that companies can implement measures that emphasize 'responsibility-by-design'. This term refers to solutions where the in-house working methods ensure that engineers have thought through the potential impact of their technology, where a responsible attitude to design is built into the workflow.

Issue:

Lack of ownership or responsibility from tech community.

Background

There is a divergence between the values the technology community sees as its responsibility in regards to AI/AS, and the broader set of social concerns raised by the public, legal, and social science communities.

The current makeup of most organizations has clear delineations between engineering, legal, and marketing arenas. Technologists feel responsible for safety issues regarding their work, but often refer larger social issues to other areas of their organization. Adherence to professional ethics is influenced by corporate values and may reflect management and corporate culture.

An organization may avoid using the word ethics, which then causes difficulties in applying generally agreed ethical standards. It is also understood that in technology or work contexts, "ethics" typically refers to a code of ethics regarding professional procedures (although codes of ethics often refer to values-driven design). Evolving language in this context is especially important as ethics regarding professional conduct often implies moral issues such as integrity or the lack thereof (in the case of whistleblowing, for instance).

Candidate Recommendations

Multidisciplinary ethics committees in engineering sciences should be generalized, and standards should be defined for how these committees operate, starting at a national level, then moving to international standards. Ethical Review Boards need to exist and to have the appropriate composition and use relevant criteria, and consider both research ethics and product

3 Methodologies to Guide Ethical Research and Design

ethics at the appropriate levels of advancement of research and development. They are not a silver bullet for all ethical conundrums, but can and should examine justifications of research or industrial projects in terms of ethical consequences. This is particularly important in the case of AI/AS as this technology is often deployed across many different sectors, politics, health care, transport, national security, the economy etc. Bringing together a multidisciplinary and diverse group of individuals will ensure that all the potential ethical issues are covered.

Further Resources

- Evolving the IRB: Building Robust Review for Industry Research by Molly Jackman of Facebook explains the differences between top down and bottom up approaches to the implementation of ethics within an organization.
- The article by van der Kloot Meijburg and ter Meulen gives a good overview of some of the issues involved in 'developing standards for institutional ethics committees'. It focuses specifically on health care institutions in the Netherlands, but the general lessons draw can also be applied to Ethical Review Boards.
- Examples of organization dealing with trade-offs (or "value trade offs") involved in the examination of the fairness of an algorithm to a specific end user population can for instance be found in the security considerations of the Internet Engineering Task Force (IETF).

Issue:

Need to include stakeholders for best context of AI/AS.

Background

Stakeholders or practitioners who will be working alongside AI and robotics technology have both interests to account for and, more importantly, insights to incorporate.

Candidate Recommendations

The interface between AI and practitioners has started to gain broader attention, e.g. IBM showing doctors using Watson,[xx] but there are many other contexts (esp. healthcare) where there may be different levels of involvement with the technology. We should recognize that, for example, occupational therapists and their assistants may have on-the-ground expertise in working with a patient, who themselves might be the "end user" of a robot or social AI technology. Their social and practical wisdom should be built upon rather than circumvented or replaced (as the dichotomy is usually framed to journalistic treatment). Technologists need to have that feedback, especially as it is not just academically oriented language about ethics but often a matter of crucial design detail gained by experience (form, sound, space, dialogue concepts).

Section 3 – Lack of Transparency

Lack of transparency about the AI/AS manufacturing process presents a challenge to ethical implementation and oversight.

Issue:

Poor documentation hinders ethical design.

Background

The limitations and assumptions of a system are often not properly documented. Oftentimes it is even unclear what data is processed or how.

Candidate Recommendations

Software engineers should be required to document all of their systems and related data flows, their performance and limitations and risks. Ethical values that have been prominent in the engineering processes should also be explicitly presented as well as empirical evidence of compliance and methodology used, such as data used to train the system, algorithms and components used and results of behavior monitoring. Criteria for such documentation could be: auditability, accessibility, meaningfulness, readability.

Further Resources

- The NATO Cybersecurity Centre for excellence (CCDCOE), addressed indicators of transparency along these lines.
- The Ethics of Information Transparency, Luciano Floridi.

Issue:

Inconsistent or lacking oversight for algorithms.

Background

The algorithms behind intelligent or autonomous systems are not subject to consistent oversight. This lack of transparency causes concern because end users have no context to know how a certain algorithm or system came to its conclusions.

Candidate Recommendations

Accountability

As touched on in the General Principles section of Ethically Aligned Design, transparency is an issue of concern. It is understood that specifics relating to algorithms or systems contain intellectual property that cannot be released to the general public. Nonetheless, standards providing oversight of the manufacturing process of intelligent and autonomous technologies need to be created to avoid harming end users.

3 Methodologies to Guide Ethical Research and Design

Michael Kearns suggests that we will need to decide to make algorithms less effective in order to achieve transparency.[xxi] Others argue that this trade-off is not necessary if we can devise new ways to ensure algorithmic accountability, for instance via the creation of an "algorithm FDA", or as suggested in a recent EU report through the creation of a regulatory body.[xxii] Although the discussion on what would be the best approach to create a standard is ongoing, the need for a standard is evident.

Policymakers are also free to restrict the scope of computational reasoning too complex to be understood in a conventional narrative or equations intelligible to humans. They may decide: if a bank can't give customers a narrative account of how it made a decision on their loan application, including the data consulted and algorithms used, then the bank cannot be eligible for (some of) the array of governmental perquisites or licenses so common in the financial field. They may even demand the use of public credit scoring models. (This is also a concern at the core of campaigns regarding lethal autonomous weapons: maybe countries should not develop killing machines powered by algorithms that evolve in unpredictable ways in response to unforeseeable stimuli).

Further Resources

- Frank Pasquale, Professor of Law at the University of Maryland, provides the following insights regarding accountability in a February, 2016 post for the Media Policy Project Blog produced by The London School of Economics and Political Science. He points out that even if machine learning processes are highly complex"*…we may still want to know what data was fed into the computational process. Presume as complex a credit scoring system as you want. I still want to know the data sets fed into it, and I don't want health data in that set—and I believe the vast majority agree with me on that. An account of the data fed into the system is not too complex for a person to understand, or for their own software to inspect. A relatively simple set of reforms could greatly increase transparency here, even if big data proxies can frustrate accountability."*

Issue:

Lack of an independent review organization.

Background

We need unaffiliated, expert opinions that provide guidance to the general public regarding automated systems and artificial intelligence. Currently, there is a gap between how AI/AS is marketed and their actual performance, or application. We need to ensure that AI/AS technology is accompanied by best use recommendations, and associated warnings. Additionally, we need to develop a certification scheme for AI/AS that ensures that the technologies have been independently assessed as being safe and ethically sound.

3 Methodologies to Guide Ethical Research and Design

For example, today it is possible for systems to download new parking intelligence to cars, and no independent reviewer establishes or characterizes boundaries or use. Or, when a companion robot like Jibo promises to watch your children, there is no organization that can issue an independent seal of approval or limitation on these devices. We need a ratings and approval system ready to serve social/automation technologies that will come online as soon as possible.

Candidate Recommendations

An independent, internationally coordinated body should be formed to oversee whether products actually meet ethical criteria, both when deployed, and considering their evolution after deployment and interaction with other products. Andrew Tutt's paper on an FDA for algorithms provides a good start. He argues that such an algorithm FDA would ensure that AI/AS develop in a way that is safe by: helping develop performance, design, and liability standards for algorithms, ensuring multi-stakeholder dialogue in the development of algorithms that are accountable and transparent, and ensure that AI/AS technology enters the market when it is deemed safe.

We also need further government funding for research into how AI/AS technologies can best be subjected to review, and how review organizations can consider both traditional health and safety issues, and ethical considerations.

Further Resources

- Tutt, Andrew. "An FDA for Algorithms." *Administrative Law Review* 67, 2016.

Issue:
Use of black-box components.

Background

Software developers regularly use 'black-box' components in their software, the functioning of which they often do not fully understand. 'Deep' machine learning processes, which are driving many advancements in autonomous systems, are a growing source of 'black-box' software. At least for the foreseeable future, AI developers will likely be unable to build systems that are guaranteed to operate exactly as intended or hoped for in every possible circumstance. Yet, the responsibility for resulting errors and harms remains with the humans that design, build, test and employ these systems.

Candidate Recommendations

When systems are built that could impact the safety or wellbeing of humans, it is not enough to just presume that a system works. Engineers must acknowledge and assess the ethical risks involved with black-box software and implement mitigation strategies where possible.

Technologists should be able to characterize what their algorithms or systems are going to do via transparent and traceable standards. To the

3 Methodologies to Guide Ethical Research and Design

degree that we can, it should be predictive, but given the nature of AI/AS systems it might need to be more retrospective and mitigation oriented.

Similar to the idea of a flight data recorder in the field of aviation, this algorithmic traceability can provide insights on what computations led to specific results ending up in questionable or dangerous behaviors. Even where such processes remain somewhat opaque, technologists should seek indirect means of validating results and detecting harms.

Software engineers should employ black-box software services or components only with extraordinary caution and ethical care, as they tend to produce results that cannot be fully inspected, validated or justified by ordinary means, and thus increase the risk of undetected or unforeseen errors, biases and harms.

Further Resources

- Pasquale, F. *The Black Box Society*. Harvard University Press, 2015.
- Another excellent resource on these issues can be found in Chava Gourarie's article, Investigating the algorithms that govern our lives (Columbia Journalism Review, April 2016). These additional recommended readings are referenced at the end of the article:
- "How big data is unfair": A layperson's guide to why big data and algorithms are inherently biased.
- "Algorithmic accountability reporting: On the investigation of black boxes": The primer on reporting on algorithms, by Nick Diakopoulos, an assistant professor at the University of Maryland who has written extensively on the intersection of journalism and algorithmic accountability.
- "Certifying and removing disparate impact": The computer scientist's guide to locating and fixing bias in algorithms computationally, by Suresh Venkatasubramanian and colleagues.
- *The Curious Journalist's Guide to Data:* Jonathan Stray's guide to thinking about data as communication, much of which applies to reporting on algorithms as well.

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 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.

As AI systems become more capable, unanticipated or unintended behavior becomes increasingly dangerous, and retrofitting safety into these more generally capable and autonomous AI systems may be difficult. Small defects in AI architecture, training, or implementation, as well as mistaken assumptions, could have a very large impact when such systems are sufficiently capable. In addition to these technical challenges, AI researchers will also confront a progressively more complex set of ethical issues during the development and deployment of these technologies.

We recommend that AI teams working to develop these systems cultivate a "safety mindset," in the conduct of research in order to identify and preempt unintended and unanticipated behaviors in their systems, and work to develop systems which are "safe by design." Furthermore, we recommend that institutions set up review boards as a resource to AI researchers and developers and to evaluate relevant projects and their progress. Finally, we recommend that the AI community encourage and promote the sharing of safety-related research and tools, and that AI researchers and developers take on the norm that future highly capable transformative AI systems "should be developed only for the benefit of all humanity and in the service of widely shared ethical ideals." (Bostrom 2014, 254) [xii]

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

Section 1 – Technical

Issue:

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.

Background

Amodei et al. (2016),[xxiv] Bostrom (2014),[xxv] Yudkowsky (2008),[xxvi] and many others have discussed how an AI system with an incorrectly or imprecisely specified objective function could behave in undesirable ways. In their paper, Concrete Problems in AI Safety, Amodei et al. describe some possible failure modes, including scenarios where the system has incentives to attempt to gain control over its reward channel, scenarios where the learning process fails to be robust to distributional shift, and scenarios where the system engages in unsafe exploration (in the reinforcement learning sense). Further, Bostrom (2012)[xxvii] and Omohundro (2008)[xxviii] have argued that sufficiently capable AI systems are likely by default to adopt "convergent instrumental subgoals" such as resource-acquisition and self-preservation, unless the objective function explicitly disincentivizes these strategies. These types of problems are likely to be more severe in systems that are more capable, unless action is taken to prevent them from arising.

Candidate Recommendation

AI research teams should be prepared to put significantly more effort into AI safety research as capabilities grow. We recommend that AI systems that are intended to have their capabilities improved to the point where the above issues begin to apply should be designed to avoid those issues pre-emptively (see the next issue stated below for related recommendations). When considering problems such as these, we recommend that AI research teams cultivate a "safety mindset" (as described by Schneier [2008][xxix] in the context of computer security), and suggest that many of these problems can likely be better understood by studying adversarial examples (as discussed by Christiano [2016] [xxx]).

We also recommend that all AI research teams seek to pursue the following goals, all of which seem likely to help avert the aforementioned problems:

1. Contribute to research on concrete problems in AI safety, such as those described by Amodei et al. in *Concrete Problems in AI Safety*[xxxi] and Taylor et al. in *Alignment for Advanced Machine Learning Systems*.[xxxii] See also the work of Hadfield-Menell et al. (2016)[xxxiii] and the references therein.

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

2. Work to ensure that AI systems are transparent, and that their reasoning processes can be understood by human operators. This likely involves both theoretical and practical research. In particular, we recommend that AI research teams develop, share, and contribute to transparency and debugging tools that make advanced AI systems easier to understand and work with; and we recommend that AI teams perform the necessary theoretical research to understand how and why a system works at least well enough to ensure that the system will avoid the above failure modes (even in the face of rapid capability gain and/or a dramatic change in context, such as when moving from a small testing environment to a large world).

3. Work to build safe and secure environments in which potentially unsafe AI systems can be developed and tested. In particular, we recommend that AI research teams develop, share, and contribute to AI safety test environments and tools and techniques for "boxing" AI systems (see Babcock et al. [2016][xxxiv] and Yampolskiy [2012][xxxv] for preliminary work).

4. Work to ensure that AI systems fail gracefully in the face of adversarial inputs, out-of-distribution errors (see Siddiqui et al. [2016][xxxvi] for an example), unexpected rapid capability gain, and other large context changes.

5. Ensure that AI systems are corrigible in the sense of Soares et al. (2015),[xxxvii] i.e., that the systems are amenable to shutdown and modification by the operators, and assist (or at least do not resist) the operators in shutting down and modifying the system (if such a task is non-trivial). See also the work of Armstrong and Orseau (2016).[xxxviii]

Issue:

Retrofitting safety into future more generally capable AI systems may be difficult.

Background

Different types of AI systems are likely to vary widely in how difficult they are to align with the interests of the operators. As an example, consider the case of natural selection, which developed an intelligent "artifact" (brains) by simple hill-climbing search. Brains are quite difficult to understand, and "refactoring" a brain to be trustworthy when given large amounts of resources and unchecked power would be quite an engineering feat. Similarly, AI systems developed by pure brute force might be quite difficult to align. At the other end of the spectrum, we can imagine AI systems that are perfectly rational and understandable. Realistic AI systems are likely to fall somewhere in between, and be built by a combination of human design and hill climbing (e.g., gradient descent, trial-and-error, etc.). Developing highly capable AI systems without these concerns in mind could result in systems with high levels of technical debt,[xl] leading to systems that are more vulnerable to the concerns raised in the previous issue stated above.

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

Candidate Recommendation

Given that some AI development methodologies will result in AI systems that are much easier to align than others, and given that it may be quite difficult to switch development methodologies late during the development of a highly capable AI system, we recommend that when AI research teams begin developing systems that are intended to eventually become highly capable, they also take great care to ensure that their development methodology will result in a system that can be easily aligned. See also the discussion of transparency tools above.

A relevant analogy for this issue is the development of the C programming language, which settled on the use of null-terminated strings [xli] instead of length-prefixed strings for reasons of memory efficiency and code elegance, thereby making the C language vulnerable to buffer overflow [xlii] attacks, which are to this day one of the most common and damaging types of software vulnerability. If the developers of C had been considering computer security (in addition to memory efficiency and code elegance), this long-lasting vulnerability could perhaps have been avoided. In light of this analogy, we recommend that AI research teams take every effort to take safety concerns into account early in the design process.

As a heuristic, when AI research teams develop potentially dangerous systems, we recommend that those systems be "safe by design," in the sense that if everything goes according to plan, then the safety precautions discussed above should not be necessary (see Christiano [2015] [xliii] for a discussion of a related concept he terms "scalable AI control"). For example, a system that has strong incentives to manipulate its operators, but which cannot due to restrictions on the system's action space, is not safe by design. Of course, we also recommend that AI research teams use all appropriate safety precautions, but safeties such as "boxes," tripwires, monitors, action limitations, and so on should be treated as fail-safes rather than as a first line of defense.

Section 2 – General Principles

Issue:

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.

Background

Issues these researchers will encounter include challenges in determining whether a system will cause unintended and unanticipated harms—to themselves, system users, and the general public—as well as complex moral and ethical considerations, including even the moral weight of certain AI systems themselves or simulations they may produce (Sandberg 2014).[xliv] Moreover, researchers are always subject to cognitive biases that might lead them to have an optimistic view of the benefits, dangers, and ethical concerns involved in their research.

Candidate Recommendation

Across a wide range of research areas in science, medicine, and social science, review boards have served as a valuable tool in ensuring that researchers are able to work with security and peace of mind about the appropriateness of their research. In addition, review boards provide a valuable function in protecting institutions, companies, and individual researchers from legal liability and reputational harm.

We recommend that organizations setup review boards to support and oversee researchers working on projects that aim to create very capable and autonomous AI systems, and that AI researchers and developers working on such projects advocate that these boards be set up (see Yampolskiy and Fox [2013][xlv] for a discussion of review boards for AI projects). In fact, some organizations like Google DeepMind and Lucid AI[xlvi] have already established review boards and we encourage others to follow their example.

Review boards should be composed of impartial experts with a diversity of relevant knowledge and experience. These boards should be continually engaged with researchers from any relevant project's inception, and events during the course of the project that trigger special review should be determined ahead of time. These types of events could include the system dramatically outperforming expectations, performing rapid self-improvement, or exhibiting a failure of corrigibility. Ideally review boards would adhere to some standards or best practices developed by the industry/field as a whole, perhaps through groups like the Partnership on Artificial Intelligence.[xlvii]

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

Given the transformative impact these systems may have on the world, it is essential that review boards take into consideration the widest possible breadth of safety and ethical issues.

Furthermore, in light of the difficulty of finding satisfactory solutions to moral dilemmas and the sheer size of the potential moral hazard that one AI research team would face when deploying a highly capable AI system, we recommend that researchers pursue AI designs that would bring about good outcomes regardless of the moral fortitude of the research team. AI research teams should work to minimize the extent to which good outcomes from the system hinge on the virtuousness of the operators.

Issue:

Future AI systems may have the capacity to impact the world on the scale of the agricultural or industrial revolutions.

Background

The development of very capable and autonomous AI systems could completely transform not only the economy, but the global political landscape. Future AI systems could bring about unprecedented levels of global prosperity, especially given the potential impact of super intelligent AI systems (in the sense of Bostrom [2014]).[xlviii] It is by no means guaranteed that this transformation will be a positive one without a concerted effort by the AI community to shape it that way (Bostrom 2014,[xlix] Yudkowsky 2008).[xlix]

Candidate Recommendations

The academic AI community has an admirable tradition of open scientific communication. Because AI development is increasingly taking place in a commercial setting, there are incentives for that openness to diminish. We recommend that the AI community work to ensure that this tradition of openness be maintained when it comes to safety research. AI researchers should be encouraged to freely discuss AI safety problems and share best practices with their peers across institutional, industry, and national boundaries.

Furthermore, we recommend that institutions encourage AI researchers, who are concerned that their lab or team is not following global cutting-edge safety best practices, to raise this to the attention of the wider AI research community without fear of retribution. Any research group working to develop capable AI systems should understand that, if successful, their technology will be considered both extremely economically significant and also potentially significant on the global political stage. Accordingly, for non-safety research and results, the case for openness is not quite so clear-cut. It is necessary to weigh the potential risks of disclosure against the benefits of openness, as discussed by Bostrom (2016).[li] Groups like the Partnership on Artificial Intelligence[lii] might help in establishing these norms and practices.

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

Finally, in his book Superintelligence, philosopher Nick Bostrom proposes that we adopt a moral norm which he calls the common good principle: "Superintelligence should be developed only for the benefit of all humanity and in the service of widely shared ethical ideals" (Bostrom 2014, 254).[liii] We encourage researchers and developers aspiring to develop these systems to take on this norm. It is imperative that the pursuit and realization of capable AI systems be done in the service of the equitable, long-term flourishing of civilization.

5 Personal Data and Individual Access Control

A key ethical dilemma regarding personal information is data asymmetry. Our personal information fundamentally informs the systems driving modern society but our data is more of an asset to others than it is to us. The artificial intelligence and autonomous systems (AI/AS) driving the algorithmic economy have widespread access to our data, yet we remain isolated from gains we could obtain from the insights derived from our lives.

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. New parameters must also be created regarding what information is gathered about individuals at the point of data collection. Future informed consent should be predicated on limited and specific exchange of data versus long-term sacrifice of informational assets.

There are a number of encouraging signs that this model of asymmetry is beginning to shift around the world. For instance, legislation like The General Data Protection Regulation (GDPR)[liv] is designed to strengthen citizens' fundamental rights in the digital age and facilitate business simplifying rules for companies by unifying regulation within the EU. Enabling individuals to curate their identity and managing the ethical implications of data use will become a market differentiator for organizations. While some may choose minimum compliance to legislation like the GDPR, forward-thinking organizations will shift their data strategy to enable methods of harnessing customer intention versus only invisibly tracking their attention.

We realize the first version of The IEEE Global Initiative's insights reflect largely Western views regarding personal data where prioritizing an individual may seem to overshadow the use of information as a communal resource. This issue is complex, as identity and personal information may pertain to single individuals, groups, or large societal data sets.

5 Personal Data and Individual Access Control

However, for any of these scenarios it is our candidate recommendation that policy should be created that:

- Allows every global citizen/individual access to tools allowing them control over a minimum common denominator of attributes that define his/her identity.
- Allows the possibility for citizens/individuals to access, manage, and control how their data is shared.
- Provides easily understandable ways for citizens/individuals to choose how or whether to share their data with other individuals, businesses, or for the common good as they choose.
- Provides for future educational programs training all citizens/individuals regarding the management of their personal data and identity, just as many countries provide training in personal finances and basic legal understanding.

We realize there are no perfect solutions, and that any digital tool can be hacked. But we need to enable a future where people control their sense of self. Augmented and virtual reality [lv] will soon provide lenses through which we perceive the world. Virtual worlds and social networks already blend our online identity with our physical sense of self. Autonomous and intelligent systems will apply virtual identities that impact the physical world.

Our goal is to champion the tools and evolved practices that could eradicate data asymmetry today to foster a positive image for our future.

5 Personal Data and Individual Access Control

Section 1 – Personal Data Definitions

The following definitions, resources, and candidate recommendations are provided to realign the systematic tracking, distribution, and storing of personal data to overtly include individuals and their predetermined preferences in the process.

Issue:

How can an individual define and organize his/her personal data in the algorithmic era?

Background

Personal data needs to embrace an individual's definition and clarification of his/her identity, mirroring unique preferences and values.

Candidate Recommendation

Where available, individuals should identify trusted identity verification resources to validate, prove, and broadcast their identity.

Further Resources

The following are two examples of identity programs along these lines:

- **eIDAS**
  Work is underway to explore extending the U.K. Verify Program to commercial applications and not just government. This aligns to the implementation of the eIDAS scheme throughout the European Union, known as Regulation (EU) N°910/2014.

  Adopted by the co-legislators in July 2014, the eIDAS scheme is a milestone to provide a predictable regulatory environment that enables secure and seamless electronic interactions between businesses, citizens, and public authorities. It ensures that people and businesses can use their own national electronic identification schemes (eIDs) to access public services in other EU countries where eIDs are available.

  The aim is to create a European internal market for eTS—namely electronic signatures, electronic seals, time stamp, electronic delivery service, and website authentication—by ensuring that they will work across borders and have the same legal status as traditional paper-based processes.

  With eIDAS, the EU has provided the foundations and a predictable legal framework for people, companies, and public administrations to safely access services and do transactions online and across borders in just "one click." Rolling out eIDAS means higher security and more convenience for any online activity such as submitting tax declarations, enrolling in a foreign university, remotely opening a bank account, setting

5 Personal Data and Individual Access Control

up a business in another Member State, or authenticating for internet payments.

- **IDNYC – New York Residents ID Program** IDNYC is a free identification card for all New York City residents. It is a government-issued photo identification card fulfilling the requirement for New York residents to permanently carry an acceptable form of ID. Eligibility extends to the most vulnerable communities; including the homeless, youth, the elderly, undocumented immigrants, the formerly incarcerated, and others who may have difficulty obtaining other government-issued ID.

  More importantly, IDNYC has implemented leading privacy practices and policies in order to further protect the vulnerable groups the program serves. These privacy enhancing processes include strict limits on the amount of time physical application documents are held before destroying them and who can access the enrollment information and the ID database, including other government and security agencies.

  Pursuant to NYC Administrative Code Section 3-115(e)(4), information collected about applicants for the IDNYC card shall be treated as confidential and may only be disclosed if authorized in writing by the individual to whom such information pertains, or if such individual is a minor or is otherwise not legally competent, by such individual's parent or legal guardian.

  The card features a photograph, name, date of birth, signature, eye color, height, and unique ID number. Residents can choose whether or not to include gender (including self-declared), emergency contact information, organ donor status, preferred language, and option to display Veteran status.

  Data privacy provides options for those that are survivors of domestic violence, or have legitimate security concerns, regarding address disclosure.

Issue:

What is the definition and scope of personally identifiable information?

Background

Personally identifiable information (PII) is defined as any data that can be reasonably linked to an individual based on their unique physical, digital, or virtual identity. As further clarification, the EU definition of personal data set forth in the Data Protection Directive 95/46/EC[lvi] defines personal data as "any information relating to an identified or identifiable natural person." The Chairwoman of the United States Federal Trade Commission has also suggested that PII should be defined broadly. The new GDPR legislation also provides

5 Personal Data and Individual Access Control

definitions for genetic and biometric data[lvii] that will become even more relevant as more devices in the Internet of Things track these unique physical identifiers.

Candidate Recommendation

PII should be considered the sovereign asset of the individual to be legally protected and prioritized universally in global, local and digital implementations. In the U.S., for instance, PII protection is often related to the right of the people to be secure in their persons, houses, papers, and effects, pursuant to the fourth amendment to the Constitution (e.g., the Supreme Court's ruling in US v. Jones from 2012, 565 U.S.).[lviii] In the EU, PII protection is commonly framed in terms of informational self-determination and defense of human dignity. In both cases, (See generally United States v. Jones, 565 U.S. 400 (2012)) the aim should be to tackle key ethical dilemmas of data asymmetry by prioritizing PII protection universally in global, local, and digital implementations.

Further Resources

- Different laws and regulations around the globe define the scope of personally identifiable information differently. The use of data analytics to derive new inferences and insights into both personal data and technical metadata raises new questions about what types of information should properly be considered personal data. This is further complicated by machine learning and autonomous systems that access and process data faster than ever before.
- The U.S. Federal Trade Commission (FTC) has taken the position in its 2009 staff report on online behavioral advertising and in its more recent 2012 Privacy Report that data is "personally identifiable," and thus warrant privacy protections, where it can be reasonably linked to a particular person, computer, or device. As a result, in many circumstances, persistent identifiers such as device identifiers, MAC addresses, static IP addresses, or cookies are considered personally identifiable under U.S. federal law. More recently, the European Court of Justice (ECJ) Advocate General has also proposed that IP addresses are personal data protected by European Union law. U.S. Federal Communications Commission (FCC) officials approved broad new privacy rules on October 27, 2016, that prevent companies like AT&T and Comcast from collecting and giving out digital information about individuals—such as the websites they visited and the apps they used— in a move that creates landmark protections for internet users. The new rules require broadband providers to obtain permission from subscribers to gather and give out data on their web browsing, app use, location, and financial information. Currently, broadband providers can track users unless those individuals tell them to stop.
- For additional discussion of how to think about what constitutes personal data, we recommend the U.K. Information Commissioner's Office paper, *Determining What Is Personal Data*, which provides guidance on how to decide whether data falls within the definition of personal data in non-obvious circumstances.

5 Personal Data and Individual Access Control

Issue:
What is the definition of control regarding personal data?

Background

Most individuals believe controlling their personal data only happens on the sites or social networks to which they belong. While taking the time to update your privacy settings on a social network is important, the logic of controlling your personal data is more holistic and universal in nature. Instead of individuals having to conform to hundreds of organization's terms and conditions or policies, in a world where people control their own personal data, those organizations would conform to an individual's predetermined requirements.

Candidate Recommendation

Personal data should be managed starting from the point of the user versus outside actors having access to data outside of a user's awareness or control.

Further Resources

- For an introduction into these issues, we recommend the Project VRM website. VRM stands for vendor relationship management, a concept created by Doc Searls and outlined with great specificity in his book, *The Intention Economy: When Customers Take Charge*. In marketing terms, customer relationship management (CRM) describes the tools utilized to track, message, and influence individuals that companies want to attract. The current Internet economy is built on this CRM model.
- Providing individuals with tools like a personal data cloud as described in the Fast Company article, "Personal.com Creates an Online Vault to Manage All Your Data," can empower users to understand how their data is an asset as well as how much data they produce. Tools like these vaults or clouds also let individuals organize their data around various uses (medical, social, banking) to potentially create an individual version of their own terms and conditions. For an example of this, we recommend reviewing Meeco.me and their Signal feature.
- For more specifics on this topic, we recommend reading Introduction to the Personal Data Ecosystem created by The Personal Data Ecosystem Consortium (PDEC).
- Hasselbalch, Gry, and Pernille Tranberg. *Data Ethics. The New Competitive Advantage*. Copenhagen: Publishare, 2016.

Section 2 – Personal Data Access and Consent

If you cannot access your personal data, you cannot benefit from its insights. Also, you will not be able to correct erroneous facts to provide the most relevant information regarding your life to the actors you trust. Multipage agreements written to protect organizations must also quickly and genuinely inform users of their choices for trusted consent in the algorithmic era.

Issue:

How can we redefine data access to honor the individual?

Background

Much of the contention associated with the concept of "privacy" actually relates to access and consent. The challenges are often around transparency and providing an explicit understanding of the consequences of agreeing to the use of our personal data, complicated by the data handling processes behind true "consent." Privacy rights are often not respected in the design and business model of services using said data.

Candidate Recommendation

Practical and implementable procedures need to be available in order for designers and developers to use "Privacy-by-Design"/Privacy-by-Default methodologies (referring to the practice or business philosophy of privacy embedded in the development of a service).

In order to realize benefits such as decision enablement and personalization for an individual, open standards and interoperability are vital to ensure individuals and society have the freedom to move across ecosystems and are not trapped by walled gardens. In order to safeguard this freedom, for example, Article 20 of the EU regulation on data protection (Right to Data Portability) sets up the right to receive PII that individuals have provided to a data controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to other controllers without hindrance from the controller to which the personal data have been provided. [lix]

Paradigms like "differential privacy" may also allow for designers and developers to bake privacy into the design and development of services.[lx] Differential privacy shifts the focus from "your data" to finding general usage

5 Personal Data and Individual Access Control

patterns across larger data sets. Differential privacy is not about anonymization of data, as that can be easily de-anonymized through intelligent cross-referencing. Instead differential privacy uses hashing, sub-sampling, and noise-injection techniques to obfuscate personal information about individuals. However, while differential privacy may provide a methodology for better usage of private or public data, it should be implemented in complement to tools and methodologies empowering individuals to manage and control their data.

As a tool for any organization regarding these issues, a good starting point is to apply the who, what, why, and when test to the collection and storage of personal information:

1. Who requires access and for what duration—is it a person, system, regulatory body, legal requirement "or" input to an algorithm?
2. What is the purpose for the access—is it read, use and discard or collect, use and store?
3. Why is the data required—is it to fulfill compliance, lower risk, because it is monetized, or in order to provide a better service/experience?
4. When will it be collected, for how long will it be kept, when will it be discarded, updated, re-authenticated—how does duration impact the quality and life of the data?

Issue:

How can we redefine consent regarding personal data so it honors the individual?

Background

Technology leaders give innovation teams and engineers too little or no direction on what human values should be considered, protected and designed for regarding personal data. When implemented correctly, solutions providing transparency and choice for the individual can be designed within the increasing regulatory environment (as is the case currently with the GDPR in the EU) to allow for minimal viable collection for maximum viable access. However, it should be noted that fundamental issues regarding the processing of data need to be addressed before exchanges happen so individuals aren't consenting to commercial or scientific usage of their information that is unclear without methods for recourse or control. A final issue to consider along these lines is how to design for portability when the convergence of digital, mobile, and Internet of Things results in the perpetual creation of data.

Candidate Recommendations

In order to realize benefits such as decision enablement and personalization for an individual, open standards and interoperability are vital to

5 Personal Data and Individual Access Control

ensure individuals and society have the freedom to move across ecosystems. Explicit consent provided by individuals in the exchange of their data via methodologies previously described in this document can inform future requirements for data to be stored, shared downstream, anonymized, or identified. By developing a decision matrix between individuals and external actors about their information, personal data can be used to process high-volume anonymized data for general insights, through to low-volume identified data used for tailored experiences.

The needs of society, communities, and research will factor into this decision matrix and introduce the need to consider security, roles, and rights management. For example, a doctor may need medical data to be identified in order to treat a patient. However a researcher may require it simply for statistical analysis and therefore does not require the data to be identifiable. Additionally mechanisms for dynamic consent as use-cases change, or data moves from the original collection context to a change of context are critical design features. This is particularly important to explicitly surface if the primary reason for data collection masks the secondary use post-collection. A European context along these lines will also require for the "right-to-be-forgotten" as a core design capability.

Further Resources

- European Commission, Factsheet on The Right to Be Forgotten Ruling.

Issue:

Data that appears trivial to share can be used to make inferences that an individual would not wish to share.

Background

How can individuals be sufficiently informed to give genuine consent?

Candidate Recommendation

While it is hoped AI/AS that parse and analyze data could also help individuals understand granular level consent in real-time, it is imperative to also put more focus on the point of data collection to minimize long-term risk.

Further Resources

As analysis becomes more autonomous, not even the analysts will necessarily know what conclusions are being drawn and used in the process. This means that informed consent could become too complex for companies to ask for or consumers to give. This is why we need to move focus away from the consent of the user to the point of data collection. Too much data is collected for no immediate purpose. There needs to be limits and exact purposes for the collection of personal data. Use limitations are also important and may be more feasible than collection limitations. Organizations should commit not to use data to make sensitive

5 Personal Data and Individual Access Control

inferences or to make important eligibility determinations.

- For an example along these lines: Felbo, B., P. Sundsøy, A. Pentland, S. Lehmann, and Y. de Montjoye. "Using Deep Learning to Predict Demographics from Mobile Phone Metadata." Cornell University Library, arXiv: 1511.06660, February 13, 2016.

Issue:

How can data handlers ensure the consequences (positive and negative) of accessing and collecting data are explicit to an individual in order for truly informed consent to be given?

Background

It is common for a consumer to consent to the sharing of discrete, apparently meaningless data points like credit card transaction data, answers to test questions, or how many steps they walk. However, once aggregated these data and their associated insights may lead to complex and sensitive conclusions being drawn about individuals that consumers would not have consented to sharing. A clear issue, as computational power increases with time and algorithms improve, is that information that was thought private can be linked to individuals at a later stage in time. Furthermore, as data is stored in terms of summaries rather than as raw observations, and may be key to training algorithms, keeping track of data usage and potential risks to privacy may be increasingly complex.

Candidate Recommendations

To guard against these types of complexities we need to make consent both conditional and dynamic. Safeguards are required to surface the downstream impact of data that appears to be trivial that can be later used to make inferences that an individual would not wish to share. Likewise, resources and legislation should be afforded to an individual so they can retract or "kill" their data if they feel it is being used in ways they do not understand or desire.

Further Resources

For examples along these lines:

- Duhigg, C. "How Companies Learn Your Secrets." *The New York Times Magazine*, Feb. 19, 2012.
- Meyer, R. "When You Fall in Love, This Is What Facebook Sees." *The Atlantic*, Feb. 15, 2014.
- Cormode, G. "The Confounding Problem of Private Data Release." *18th International Conference on Database Theory* (2015): 1–12. DOI: 10.4230/LIPIcs.ICDT.2015.1.

Section 3 – Personal Data Management

For individuals to achieve and retain a parity regarding their personal information in the algorithmic age, it will be necessary to extend an Identity Assurance paradigm to include a proactive algorithmic tool that acts as their agent or guardian in the digital, and "real" world ("real" meaning a physical or public space where the user is not aware of being under surveillance by facial recognition, biometric, or other tools that could track, store, and utilize their data without pre-established consent or permission).

Issue:

Could a person have a personalized AI or algorithmic guardian?

Background

The creation of a personalized AI would provide a massive opportunity for innovation in AI and corporate communities. Some might view an individual's desire to control and manage their data as hindering innovation since higher choices may conflict with well-intentioned efforts to amass vast data sets for public good. However, this view inherently assumes all individuals in a certain context would want their data utilized for a certain project, even if it was for the "public good."

The sophistication of data-sharing methodologies have evolved so these scenarios can evolve from an "either/or" relationship ("we get all of your data for this project or you provide nothing and hinder this work") to a "yes and" one—by allowing individuals to set their preferences for sharing and storing their data they are more likely to trust the organizations conducting research and provide more access to their data.

It should also be noted that providing these types of platforms and paradigms is of value to organizations at large because contrary to rhetoric saying, "privacy is dead," individuals and governments around the world have become more focused on the control of privacy and personal data in the past few years. In the United States, according to a May 20, 2015 report, "93% of adults say that being in control of *who* can get information about them is important; 74% feel this is 'very important,' while 19% say it is 'somewhat important'" and, "90% say that controlling what information is collected about them is important—65% think it is 'very important' and 25% say it is 'somewhat important' (Madden and Rainie)."[lxi]

5 Personal Data and Individual Access Control

Candidate Recommendation

Algorithmic guardian platforms should be developed for individuals to curate and share their personal data. Such guardians could provide personal information control to users by helping them track what they have agreed to share and what that means to them while also scanning each user's environment to set personal privacy settings accordingly. The guardian could serve as an educator and negotiator on behalf of its user by suggesting how requested data could be combined with other data that has already been provided, inform the user if data is being used in a way that was not authorized, or make recommendations to the user based on a personal profile. As a negotiator, the guardian could negotiate conditions for sharing data and could include payment to the user as a term, or even retract consent for the use of data previously authorized for a breach of conditions.

Nonetheless, the dominant paradigm for personal data models needs to shift to being person-based and away from system and service-based models not under the control of the individual/human. Personal data cannot be controlled or understood when fragmented and controlled by a myriad of entities in legal jurisdictions across the world. The object model for personal data should be associated with that person, and under the control of that person utilizing a personalized AI or algorithmic guardian. *Specifically*:

- For purposes of privacy, a person must be able to set up any number of agents/guardians or profiles within one agent with different levels or types of personal data associated.
- During the handshake/negotiation between the personal agent and the system or service, if the required data set contains elements the personal agent will not provide, the service may be unavailable. If the recommended data set will not be provided, the service may be degraded.
- Default profiles, to protect naive or uninformed users, should provide little or no personal information without explicit action by the personal agent's owner.

Further Resources

- We wish to acknowledge Jarno M. Koponen's articles on Algorithmic Angels that provided inspiration for portions of these ideas.
- Companies are already providing solutions for early or partial versions of algorithmic guardians. Anonyome Labs recently announced their SudoApp that leverages strong anonymity and avatar identities to allow users to call, message, email, shop, and pay—safely, securely, and privately.
- Tools allowing an individual to create a form of an algorithmic guardian are often labeled as PIMS, or personal information management services. Nesta in the United Kingdom was one of the funders of early research about PIMS conducted by CtrlShift.

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 are not designed to cause harm. Multi-year discussions on international agreements around autonomous systems in the context of war are occurring at the UN, but professional ethics about such systems 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 the implications of their work, and that professional ethical codes appropriately address works that are intended to cause harm.

Specifically, we would like to ensure that stakeholders are working with sensible and comprehensive shared definitions of concepts relevant in the space of autonomous weapons systems (AWS). We recommend designers not only take stands to ensure meaningful human control, but be proactive about providing quality situational awareness through those autonomous or semi-autonomous systems to the humans using those systems. Stakeholders must recognize that the chains of accountability backward, and predictability forward, also include technical aspects such as verification and validation of systems, as well as interpretability and explainability of the automated decision-making, both in the moment and after the fact.

A concern is that professional ethical codes should be informed by not only the law but an understanding of both direct and macro-level ramifications of products and solutions developed explicitly as, or that can be expected to be used or abused as, AWS. Some types of AWS are particularly societally dangerous because they are too small, insidious, or obfuscated to be attributable to the deploying entity, and so ethical recommendations are needed to prevent these instances from having dangerous outcomes.

6 Reframing Autonomous Weapons Systems

Issue:

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.

Background

Many professional organizations have codes of conduct intended to align individuals' behaviors toward particular values; however, they seldom sufficiently address members' behaviors in contributing toward particular artifacts, such as creating technological innovations deemed threatening to humanity, especially when those innovations have significant probabilities of costly outcomes to people and society. Foremost among these in our view are technologies related to the design, development, and engineering of AWS.

Candidate Recommendations

- We propose that any code of conduct be extended to govern a member's choice to create or contribute to the creation of technological innovations that are deemed threatening to humanity. Such technologies carry with them a significant probability of costly outcomes to people and society. When codes of conduct are directed towards ensuring positive benefits or outcomes for humanity, organizations should ensure that members do not create technologies that undermine or negate such benefits. In cases where created technologies or artifacts fail to embody or conflict with the values espoused in a code of conduct, it is imperative that professional organizations extend their codes of conduct to govern these instances so members have established recourse to address their individual concerns. We also recommend that codes of conduct more broadly ensure that the artifacts and agents offered into the world by members actively reflect the professional organization's standards of professional ethics.
- Professional organizations need to have resources for their members to make inquiries concerning whether a member's work contravenes International Humanitarian Law or International Human Rights Law.

Further Resources

- Kvalnes, Øyvind. "Loophole Ethics," in *Moral Reasoning at Work: Rethinking Ethics in Organizations*, 55–61. Palgrave Macmillan U.K., 2015.
- Noorman, Merel. "Computing and Moral Responsibility," *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), Summer 2014 Edition.
- Hennessey, Meghan. "ClearPath Robotics Takes Stance Against 'Killer Robots'." ClearPath Robotics, 2014.
- "Autonomous Weapons: An Open Letter from AI & Robotics Researchers." Future of Life Institute, 2015.

6 Reframing Autonomous Weapons Systems

Issue:

Confusions about definitions regarding important concepts in artificial intelligence (AI), autonomous systems (AS), and autonomous weapons systems (AWS) stymie more substantive discussions about crucial issues.

Background

The potential for confusion about definitions is not just an academic concern. The lack of clear definitions regarding AWS is often cited as a reason for not proceeding toward any kind of international control over autonomous weapons.

The term autonomy is important for understanding debates about AWS; yet there may be disputes—about what the term means and whether it is currently possible—that prevent progress in developing appropriate policies to guide its design and manufacture. We need consistent and standardized definitions to enable effective discussions of AWS, free from technological considerations that are likely to be quickly outdated. As this is both a humanitarian issue and an issue of geopolitical stability, the focus in this area needs to be on how the weapons are controlled by humans rather than about the weapons technology per se.

The phrase "in the loop" also requires similar clarification. Let us assume that an automatic weapons system requests permission to fire from a human operator, and the operator gives permission. How long of a delay should be acceptable between the system request and the operator's permission take place before the situation has changed to invalidate the permission? A sub-second clearance would probably be judged as acceptable in most cases, but what about multiple minutes? It could be argued that the situation itself should be examined, but that may result in either undue cognitive load on the operator at a critical time, or for the system itself to make decisions on what is "an appropriate level of change" and possibly retract its intent to fire.

What is often also unclear in these scenarios is whether clearance to fire at a target means a system is cleared to prosecute that target indefinitely, or has predetermined limits on the amount of time or ordinance each clearance provides.

In analyzing these issues, one quickly realizes that the type of autonomy that is of concern is no more complicated than the type of autonomy that we cede to chess programs. In both cases the human has not anticipated in advance and made an appropriate decision for every situation that can possibly arise. In many cases the machine's decision in these instances will be different from what the human's decision would have been.

This notion of autonomy can be applied separately to each of the many functions of a weapons system; thus, an automatic weapons system could be autonomous in searching

6 Reframing Autonomous Weapons Systems

for targets but not in choosing which ones to attack, or vice versa. It may or may not be given autonomy to fire in self-defense when the program determines that the platform is under attack, and so on. Within each of these categories, there are also many intermediate gradations in the way that human and machine decision making may be coupled.

Candidate Recommendations

- The term *autonomy* in the context of AWS should be understood and used in the restricted sense of delegation of decision-making capabilities to a machine. Since different functions within AWS may be delegated to varying extents, and the consequences of such delegation depend on the ability of human operators to forestall negative consequences via the decisions over which they retain effective control, it is important to be precise about the ways in which control is shared between human operators and AWS.
- We recommend that various authorization scenarios be further investigated for ethical best practices by a joint workshop of stakeholders and concerned parties (including, but not limited to, international humanitarian organizations and militaries), and that those best practices be promoted by professional organizations as well as by international law.

Further Resources

- Dworkin, Gerald. *The Theory and Practice of Autonomy*. Cambridge University Press, 1988.
- Frankfurt, Harry G. "Freedom of the Will and the Concept of a Person," in T*he Importance of What We Care About*, Cambridge University Press, 1987.
- DoD Defense Science Board, The Role of Autonomy in DoD Systems, Task Force Report, July 2012, 48.
- DoD Defense Science Board, Summer Study on Autonomy. June 2016.
- Young, Robert. Autonomy: *Beyond Negative and Positive Liberty*. St. Martin's Press, 1986.
- Society of Automotive Engineers standard J3016, Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems, 2014.
- Sheridan, T. B., and W. L. Verplank. Human and Computer Control of Undersea Teleoperators.. Cambridge, MA: Man-Machine Systems Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology, 1978.
- Sharkey, Noel. "Towards a Principle for the Human Supervisory Control of Robot Weapons." *Politica and Società* 2 (2014): 305–324.

6 Reframing Autonomous Weapons Systems

Issue:
AWS are by default amenable to covert and non-attributable use.

Background

The lack of a clear owner of a given AWS incentivizes scalable covert or non-attributable uses of force by state and non-state actors. Such dynamics can easily lead to unaccountable violence and societal havoc.

Candidate Recommendation

Because AWS are delegated authority to use force in a particular situation, they are required to be attributable to the entity that deployed them through the use of physical external and internal markings as well as within their software.

Further Resources

- Bahr, Elizabeth. "Attribution of Biological Weapons Use," in *Encyclopedia of Bioterrorism Defense*. John Wiley & Sons, 2005.
- Mistral Solutions. "Close-In Covert Autonomous Disposable Aircraft (CICADA) for Homeland Security,"2014.
- Piore, Adam. "Rise of the Insect Drones." Wired. January 29, 2014.

Issue:
There are multiple ways in which accountability for AWS's actions can be compromised.

Background

Weapons may not have transparency, auditability, verification, or validation in their design or use. Various loci of accountability include those for commanders (e.g., what are the reasonable standards for commanders to utilize AWS?), and operators (e.g., what are the levels of understanding required by operators to have knowledge of the system state, operational context, and situational awareness?).

There are currently weapons systems in use that, once activated, automatically intercept high-speed inanimate objects such as incoming missiles, artillery shells, and mortar grenades. Examples include C-RAM, Phalanx, NBS Mantis, and Iron Dome. These systems complete their detection, evaluation, and response process within a matter of seconds and thus render it extremely difficult for human operators to exercise meaningful supervisory control once they have been activated other than deciding when to switch them off. This is called supervised autonomy by the US DoD[lxii] because the weapons require constant and vigilant human evaluation and monitoring for rapid shutdown in cases of targeting errors, change of situation, or change in status of targets.

6 Reframing Autonomous Weapons Systems

Candidate Recommendations

- Trusted user authentication logs and audit trail logs are necessary, in conjunction with meaningful human control. Thorough factors-driven design of user interface and human–computer/robot interaction design is necessary for situational awareness, knowability, understandability and interrogation of system goals, reasons and constraints, such that the user could be held culpable.
- Tamper-proof the equipment used to store authorization signals and base this on open, auditable designs, as suggested by Gubrud and Altmann (2013). Further, the hardware that implements the human-in-the-loop requirement should not be physically distinct from the operational hardware of the system, to deter the easy modification of the overall weapon after the fact to operate in fully autonomous mode.
- System engineers must have higher standards and regulations of security for system design from a cybersecurity perspective than they would for other computer-controlled weapons systems. AWS ought to be designed with cybersecurity in mind such that preventing tampering, or at least undetected tampering, is a highly weighted design constraint.

Further Resources

- Gubrud, M., and J. Altmann. "Compliance Measures for an Autonomous Weapons Convention." International Committee for Robot Arms Control, May 2013.
- The UK Approach to Unmanned Aircraft Systems (UAS), Joint Doctrine Note 2/11, March 30, 2011.
- Sharkey, Noel. "Towards a Principle for the Human Supervisory Control of Robot Weapons." *Politica and Società* 2 (2014): 305–324.
- Owens, D. "Figuring Forseeability." *Wake Forest Law Review* 44 (2009): 1277, 1281–1290.
- Scherer, Matt. "Who's to Blame (Part 4): Who's to Blame if an Autonomous Weapon Breaks the Law?" *Law and AI* (blog), February 24, 2016.

Issue:

An automated weapons system might not be predictable (depending upon its design and operational use). Learning systems compound the problem of predictable use.

Background

Modeling and simulation of AWS, particularly learning systems, may not capture all possible circumstances of use or situational interaction. They are underconstrained cyberphysical systems. Intrinsic unpredictability of adaptive systems is also an issue: one cannot accurately model one's adversary's systems and how an

6 Reframing Autonomous Weapons Systems

adversary will adapt to your system resulting in an inherently unpredictable act.

Candidate Recommendation

The predictability of the overall user-system-environment combination should be striven for. Having a well-informed human in the loop will help alleviate issues that come with open-world models and should be mandated.

Further Resources

- International Committee for Robot Arms Control. "LAWS: Ten Problems for Global Security" (leaflet). 10 April 2015.
- Owens, D. "Figuring Forseeability." *Wake Forest Law Review* 44 (2009): 1277, 1281–1290.
- Scherer, Matt. "Who's to Blame (Part 5): A Deeper Look at Predicting the Actions of Autonomous Weapons." *Law and AI* (blog), February 29, 2016.

Issue:

Legitimizing AWS development sets precedents that are geopolitically dangerous in the medium-term.

Background:

The widespread adoption of AWS by major powers would destabilize the international security situation by:

A. Allowing an autonomous weapon to initiate attacks in response to perceived threats, leading to unintended military escalation or war;

B. Creating weapons that adapt their behavior to avoid predictability, thereby reducing humans' ability to foresee the consequences of deployment;

C. Creating a fragile strategic balance that depends largely on the capabilities of autonomous weapons, which can change overnight due to software upgrades or cyber-infiltration; and,

D. Allowing the dynamics of constant incursions, similar to those faced in the cyberwarfare sphere, where offense is asymmetrically easier than defense, to enter the kinetic sphere.

Due to the iterative and competitive nature of weapons acquisition and use, the development and deployment of AWS creates incentives for further use and development of more sophisticated AWS in all domains. This cycle incentivizes faster decision-making in critical situations and conflicts, and more complex and less scrutable or observable processes, thereby excluding human participation in decision-

6 Reframing Autonomous Weapons Systems

making. Hence, by decoupling the number of weapons that can be deployed in an attack from the number of humans required to manage the deployment, AWS lead to the possibility of scalable weapons of mass destruction whose impact on humanity is likely to be negative.

AWS use will likely give rise to rapid escalation of conflict due to their purpose of increasing operational efficiency and tempo. Thus, there will likely be little or no opportunity for human commanders to deliberate and perform de-escalation measures in scenarios where such weapons are deployed on multiple sides of a conflict or potential conflict.

Use of AWS by two parties (or more) to a conflict will likely lead to complex interactions that are difficult to model, understand, and control. AWS also enable oppression through suppression of human rights, both in domestic and international settings, by enabling new scalabilities in enacting potentially illegal or unethical orders that human soldiers might reject.

AWS's ability to decouple the number of weapons that can be deployed in an attack from the number of humans required to manage their deployment leads to the possibility of scalable weapons of mass destruction whose impact on humanity is likely to be negative.

There is, thus, a dual, and interactive concern with regards to AWS:

1. The nature of inter-state competition in arms races yields escalatory effects with regards to arms development, deployment and proliferation; and
2. The very nature of AI in competitive and cyclical environments drives toward goal-maximizing behavior that without sufficient safeguards enables "flash crash"–type scenarios.[lxiii]

Candidate Recommendation

Autonomy in functions such as target selection, attack, and self-defense leads to negative consequences for humanity, and therefore should be curtailed by designing systems which require human involvement in such decisions. There must be meaningful human control over individual attacks.

Design, development, or engineering of AWS beyond meaningful human control that is expected to be used offensively or kill humans is to be unethical. Such systems created to act outside of the boundaries of "appropriate human judgment," "effective human control," or "meaningful human control," undermine core values technologists adopt in their typical codes of conduct.

Further Resources

- Scharre, P., and K. Sayler. "Autonomous Weapons and Human Control" (poster). Center for a New American Security, April 2016.
- International Committee for Robot Arms Control. "LAWS: Ten Problems for Global Security" (leaflet). April 10, 2015.

6 Reframing Autonomous Weapons Systems

Issue:

Exclusion of human oversight from the battlespace can too easily lead to inadvertent violation of human rights and inadvertent escalation of tensions.

Background

The ethical disintermediation afforded by AWS encourages the bypassing of ethical constraints on people's actions that should require the consent of multiple people, organizations, or chains of commands. This exclusion concentrates ethical decision making into fewer hands

Candidate Recommendation:

Design, development, or engineering of AWS for anti-personnel or anti-civilian use or purposes are unethical. An organization's values on respect and the avoidance of harm to persons precludes the creation of AWS that target human beings. If a system is designed for use against humans, such systems must be designed as semi-autonomous where the control over the critical functions remains with a human operator, (such as through a human-in-the-loop hardware interlock). Design for operator intervention must be sensitive to human factors and increasing—rather than decreasing—situational awareness. Under no circumstances is it morally permissible to use predictive or anticipatory AWS against humans. "Preventive self-defense" is not a moral justification in the case of AWS.
Ultimately, weapons systems must be under meaningful human control. AWS operating without meaningful human control should be prohibited, and as such design decisions regarding human control must be made so that a commander has meaningful human control over direct attacks during the conduct of hostilities. In short, this requires that a human commander be present and situationally aware of the circumstances on the ground as they unfold to deploy either semi-autonomous or defensive anti-materiel AWS. Organizational members must ensure that the technologies they create enhance meaningful human control over increasingly sophisticated systems and do not undermine or eliminate the values of respect, humanity, fairness, and dignity.

Further Resources

- International Committee for Robot Arms Control. "LAWS: Ten Problems for Global Security" (leaflet), April 10, 2015.
- Heller, Kevin Jon. "Why Preventive Self-Defense Violates the UN Charter." *Opinio Juris* (blog), March 7, 2012.
- Scherer, Matt. "Who's to Blame (Part 5): A Deeper Look at Predicting the Actions of Autonomous Weapons." *Law and AI* (blog), February 29, 2016.

6 Reframing Autonomous Weapons Systems

Issue:

The variety of direct and indirect customers of AWS will lead to a complex and troubling landscape of proliferation and abuse.

Background

Use of AWS by a myriad of actors of different kinds, including states (of different types of regime) and non-state actors (militia, rebel groups, individuals, companies, including private military contractors) would lead to such systems becoming commonplace anywhere anyone favors violence due to the disintermediation and scalability afforded by their availability.

There will be incentives for misuse depending upon state of conflict and type of actor. For example, such misuse may include, but is not limited to, political oppression, crimes against humanity, intimidation, assassination, and terrorism. This can lead to, for example, a single warlord targeting an opposing tribe based on their respective interests as declared on Facebook, their DNA, their mobile phones, or their looks.

Candidate Recommendations

- There is an obligation to know one's customer. One must design AWS in such a way that avoids tampering for unintended use. Further work on technical means for nonproliferation should be explored, for example, cryptographic chain authorization.
- There is an obligation to consider the foreseeable use of the system, and whether there is a high risk for misuse.
- There is an obligation to consider, reflect on, or discuss possible ethical consequences of one's research and/or the publication of that research.

Issue:

By default, the type of automation in AWS encourage rapid escalation of conflicts.

Background

One of the main advantages cited regarding autonomous weapons is that they can make decisions faster than humans can, enabling rapid defensive and offensive actions. When opposing autonomous weapons interact with each other, conflict will be able to escalate more quickly than humans on either side will be able to understand.

Candidate Recommendation

- Consider ways of limiting potential harm, for example, limited magazines, munitions, or maximum numbers of platforms in collaborative teams. Explore other technological means for limiting escalation, for example, "circuit breakers," as well as features that can support confidence-building measures between adversaries, for example, methods to communicate. All such

6 Reframing Autonomous Weapons Systems

solution options ought to precede the design, development, deployment, and use of AWS.

- Perform further research on how to temper such dynamics when designing these systems.

Issue:

There are no standards for design assurance verification of AWS.

Background

Standards for guaranteeing the compliance of autonomous and semi-autonomous weapons systems with relevant ethical and legal standards are lacking. Comprehensive international standards are needed to ensure this complex topic receives the critical evaluative process it merits.

Candidate Recommendation

It should be feasible to discern and verify that a system meets the relevant ethical and legal standards, such as international humanitarian law. We recommend efforts to standardize a comprehensive suite of verification and validation protocols for AWS and semi-autonomous weapons. Stakeholders including humanitarian organizations and AI safety concerns should contribute to the technical requirements for this.

Further Resources

- International Standards Organization. ISO 13849-1:2015: Safety of Machinery—Safety-Related Parts of Control Systems, General Principles for Design.

Issue:

Understanding the ethical boundaries of work on AWS and semi-autonomous weapons systems can be confusing.

Background

While national laws may differ on what constitutes responsibility or liability for the design of a weapons' system, given the level of complicity or the causal contribution to the development of a technology, ethics looks for lines of moral responsibility. Determining whether one is morally responsible requires us to establish relevant facts in relation to a person's acts and intentions.

Candidate Recommendation

How one determines the line between ethical and unethical work on AWS requires that one address whether the development, design, production, and use of the system under consideration is itself ethical. It is incumbent upon a member to engage in reflective judgment to consider whether or not his or

6 Reframing Autonomous Weapons Systems

her contribution will enable or give rise to AWS and their use cases. Members must be aware of the rapid, dynamic, and often escalatory natures of interactions between near-peer geopolitical adversaries or rivals. It is also incumbent upon members of a relevant technical organization to take all reasonable measures to inform themselves of the funding streams, the intended use or purpose of a technology, and the foreseeable misuse of their technology when their contribution is toward AWS in whole or in part. If their contribution to a system is foreseeably and knowingly to aid in human-aided decisions—that is, as part of a semi-autonomous weapons system—this may act as a justification for their research.

Further Resources

- Sharkey, N. "Cassandra or the False Prophet of Doom: AI Robots and War." *IEEE Intelligent Systems* 28, no. 4 (2008): 14–17.
- Noorman, Merel. "Computing and Moral Responsibility," in *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), edited by Edward N. Zalta.

7 Economics/Humanitarian Issues

It is irrefutable that 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 public feels unprepared personally and professionally in a period of dramatic change. Overly optimistic advocacy about the positive outcomes competes with legitimate concerns on the emerging individual and institutional harms related to privacy, discrimination, equity, security of critical infrastructure, and other issues. Dialogue about the effects of technology on people is needed with respect to those technologies that can have a longer term, chronic effect on human wellbeing. A more balanced, granular, analytical, and objective treatment of this subject will more effectively help inform policy making, and has been sorely lacking to date. A concerted effort is required between and among technologists, ethicists, civil society, and public policymakers on how to identify and measure gaps, barriers, and benefits, and to initiate a sustainable, scalable dialogue between and among different stakeholders.

As part of our "systems-engineering" approach to human-technology systems, emphasis has been placed on approaches (such as shared metrics, taxonomy conversion tables, hybrid and integrated incentives and penalty structures, etc.) that can best integrate the learning about human and social wellbeing from a number of perspectives such as environmental, cultural, political, socio-economic, and resource constraints. Also, the "system" scope at issue is considered to include the encounters between information-fueled technologies and the entire human species. This scope, in turn, invites an analytical construction of problems and potential solutions that can address both current issues in developed countries and also humanitarian issues in developing economies; recognizing that some problems and solutions will be more "localized" within a given population than others.

The aim of our multi-stakeholder Committee is to identify the key drivers shaping the human-technology global ecosystem, and to suggest key opportunities for solutions that could be implemented by unlocking critical choke points of tension. In fact, the presence of various "tensions" viewed from economic, social, cultural, political, and other perspectives provide signposts of entrepreneurial opportunity—each is an opportunity to address perceived "risk arbitrage" of multiple parties—with the potential for generating great value

7 Economics/Humanitarian Issues

from holistic solutions. With this shared understanding of the power dynamics across various categories of stakeholders, our goal is to create the beginnings of a shared agenda with a prioritized set of actions. 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.

7 Economics/Humanitarian Issues

Section 1 – Automation and Employment

While there is evidence that robots and automation are taking jobs away in various sectors, a more balanced, granular, analytical, and objective treatment of this subject will more effectively help inform policy making, and has been sorely lacking to date.

Issue:

Misinterpretation of artificial intelligence and autonomous systems (AI/AS) in media is confusing to the public.

Background

Information, analysis, and disinformation in the media regarding robotics/AI and jobs tend to focus on gross oversimplifications such as doom and utopia. This does not help in starting an objective debate and sends a wrong message to the general public.

Candidate Recommendation

Create an international, independent information clearinghouse that can properly disseminate objective statistics, fact-check and generally inform media, policymakers, the general public and other stakeholders about the impact of robotics and AI on jobs, growth, and new employment structures.

Issue:

Automation is not typically viewed only within market contexts.

Background

Robotics and AI are expected to have an impact beyond market domains and business models. Examples of impact include safety, public health, and socio-political considerations of deploying robotics/AI systems. This impact will diffuse through the global society.

Candidate Recommendation

In order to properly understand the impact of robotics/AI on society including those related to employment, it is necessary to consider both product and process innovation as well as wider implications from a global perspective.

7 Economics/Humanitarian Issues

Further Resources

- Pianta, M. Innovation and Employment, Handbook of Innovation. Oxford University Press, 2003.
- Vivarelli, M. Innovation and Employment: A Survey, Institute for the Study of Labor (IZA) Discussion Paper No. 2621, 2007.

Issue:

The complexities of employment are being neglected regarding robotics/AI.

Background

Current attention on automation and employment tends to focus on the sheer number of jobs lost or gained. Other concerns include changes in the traditional employment structure(s).

Candidate Recommendation

It is important to focus the analysis on how the structures surrounding employment structure will be changed by automation and AI rather than on solely dwelling on the number of jobs that might be impacted. The analysis should focus on how current task content of jobs are changed based on a clear assessment of the automatibility of the occupational description of such jobs.

Further Resources

- RockEU. Robotics Coordination Action for Europe Report on Robotics and Employment.

Issue:

Technological change is happening too fast for existing methods of (re)training the workforce.

Background

The current pace of technological change would heavily influence changes in the employment structure. In order to properly prepare the workforce for such evolution, actions should be proactive and not only reactive.

Candidate Recommendations

To cope with the technological pace and ensuing progress, it will be necessary that workers improve their adaptability to rapid technological changes through adequate training programs provided to develop appropriate skillsets. Training programs should be available to any worker with special attention to the low-skilled workforce. Those programs can be private (sponsored by the employer) or public (offered freely through specific public channels and policies), and they should be open while the worker is in-between jobs or still employed. Fallback strategies also need to be developed for those who cannot be re-trained.

7 Economics/Humanitarian Issues

Issue:
AI policy may slow innovation.

Background

There exists a false concern that policy and regulation necessarily slows down innovation. However, it is important that emerging technologies should be regulated such that their adverse effects on society are minimized. This requires agility in governance.

Candidate Recommendation

It is imperative that legislation and AI policy are nimble enough to keep up with the rapid advancement of technology while proposing rules and regulations that protect societal values and facilitate, rather than unnecessarily stymie, innovation. Close collaboration of governments, industries, and civil society take on a renewed meaning more than ever, given these concerns.

Section 2 – Accountability and Equal Distribution

For AI systems to be adopted in an atmosphere of trust and safety, greater efforts must be undertaken to increase transparency, clarity, and availability of these resources.

Issue:

AI and autonomous technologies are not equally available worldwide.

Background

We need to ensure the equitable distribution of the benefits of AI/AS technology worldwide. Training, education, and opportunities in robotics and autonomous systems worldwide should be provided particularly with respect to underdeveloped nations.

Candidate Recommendation

Working with appropriate organizations (e.g., United Nations, OAS, etc.) stakeholders from a cross-sectional combination of government, corporate, and NGO communities should:

1. Engage in discussions regarding effective education and training;
2. Encourage global standardization/ harmonization and open source software; and,
3. Promote distribution of knowledge and wealth generated by the latest autonomous systems, including formal financial mechanisms (such as taxation or donations to effect such equity worldwide).

Issue:

Lack of access and understanding regarding personal information.

Background

How to handle privacy and safety issues, especially as it applies to data in humanitarian and development contexts?

Candidate Recommendation

Urgent issues around individual consent, potential privacy breaches, and potential for harm or discrimination regarding individual's personal data require attention and standardized approaches. This is especially true with populations that are recently online, or lacking a good understanding of data use and "ownership," privacy, and how their digital access generates personal data by-products used by third parties.

7 Economics/Humanitarian Issues

According to GSMA, the number of mobile Internet users in the developing world will double from 1.5 billion in 2013 to 3 billion by 2020, rising from 25 percent of the developing world population to 45 percent over the period. [lxiv] In Sub-Saharan Africa, just 17 percent of the population were mobile Internet subscribers in 2013, but penetration is forecast to increase to 37 percent by 2020—making the generation, storage, use, and sharing of personal data in the developing world an issue that will continue to gain gravity.

In the humanitarian sector, digital technologies have streamlined data collection and data sharing, frequently enabling improved outcomes. With a focus on rights and dignity of the populations served, practitioners and agencies have advocated for more data sharing and open data in the social good sector. Timely access to public, social sector, and private data will speed response, avoid collection duplications, and provide a more comprehensive summary of a situation, based on multiple data streams and a wider range of indicators.

However, there are inherent risks when multiple sources of data are overlaid and combined to gain insights, as vulnerable groups or individuals can be inadvertently identified in the process. The privacy threat is the most discussed risk: When is informed consent or opt-in really ethical and effective? Best practices remain an unresolved issue among practitioners when working with communities with fewer resources, low literacy, lower connectivity, and less understanding about digital privacy.

The "do no harm" principle is practiced in emergency and conflict situations. Humanitarian responders have a responsibility to educate the populations about what will happen with their data in general, and what might happen if it is shared openly; there is often lack of clarity around how these decisions are currently being made and by whom. Remedial steps should include community education regarding digital privacy, as well as helping vulnerable groups become more savvy digital citizens.

There are perception gaps regarding what constitutes potential and actual harm stemming from data use practices. A collaborative consensus across sectors is needed on safeguarding against risks in data collection, sharing, and analysis—particularly of combined sets. From the outset, iterative, ethics-based approaches addressing data risk and privacy are key to identify and mitigate risks, informing better action and decision-making in the process.

Further Resources

- For more on responsible data use, see the Responssible Development Data Book. Oxfam also has a responsible data policy that provides a field-tested reference.
- Example Use Case from GSMA: When Call Data Records (CDRs) are used to help in the response to the Ebola outbreak, mobile operators wish to ensure mobile users' privacy is respected and protected and associated risks are addressed.

Section 3 – Empowering Developing Nations to Benefit from AI

Many of the debates surrounding AI/AS take place within advanced countries among individuals benefiting from adequate finances and higher than average living situations. It is imperative that all humans in any condition around the world are considered in the general development and application of these systems to avoid the risk of bias, classism, and general non-acceptance of these technologies.

Issue:

Increase the active representation of developing nations in The IEEE Global Initiative.

Background

At the point of its first release, The Global Initiative is primarily made up of individuals from North America and Europe.

Candidate Recommendation

Representatives from developing countries should be part of every committee of *Ethically Aligned Design* so that proper concerns are accurately reflected. The conditions that would facilitate the inclusion of inputs from developing nations should be fostered.

Institute educational initiatives for universities, industry, and government to promote a balanced understanding of robotics/AI risks, benefits, and consequences. Scholarships, exchange programs, and distinguished lecturer series are some possible ways this can be realized.

Issue:

The advent of AI/AS can exacerbate the economic and power structure differences between and within developed and developing nations.

Background

How will developing nations implement AI/AS via existing resources? Do the economics of developing nations allow for AI/AS implementation? How can people without technical expertise maintain these systems?

7 Economics/Humanitarian Issues

Candidate Recommendation

Develop mechanisms for increasing transparency of power structures and justly sharing the economic and knowledge acquisition benefits of robotics/AI. Facilitate robotics/AI research and development in developing nations. Ensure that representatives of developing nations are involved.

Further Resources

- Ajakaiye, O., and M. S. Kimenyi. "Higher Education and Economic Development in Africa: Introduction and Overview." *Journal of African Economies* 20, no. 3 (2011): iii3–iii13.
- Bloom, D. E., D. Canning, and K. Chan. *Higher Education and Economic Development in Africa* (Vol. 102). Washington, DC: World Bank, 2006.
- Brynjolfsson, E., and A. McAfee. The *Second Age of Machine Intelligence: Work Progress and Prosperity in a Time of Brilliant Technologies*. New York: W. W. Norton & Company, 2014.
- Dahlman, C. Technology, *Globalization, and Competitiveness: Challenges for Developing Countries. Industrialization in the 21st Century.* New York: United Nations, 2006.
- Fong, M. Technology *Leapfrogging for Developing Countries. Encyclopedia of Information Science and Technology*, 2nd ed. Hershey, PA: IGI Global, 2009 (pp. 3707–3713).
- Frey, C. B., and M. A. Osborne. "The Future of Employment: How Susceptible Are Jobs to Computerisation?" (working paper). Oxford University, 2013.
- Rotman, D. "How Technology Is Destroying Jobs." *MIT Technology Review*, June 12, 2013.
- McKinsey Global Institute. "Disruptive Technologies: Advances That Will Transform Life, Business, and the Global Economy" (report), May 2013.
- Sauter, R., and J. Watson. "Technology Leapfrogging: A Review of the Evidence, A Report for DFID." Brighton, England: University of Sussex. October 3, 2008.
- *The Economist.* "Wealth Without Workers, Workers Without Wealth." October 4, 2014. .
- World Bank. "Global Economic Prospects 2008: Technology Diffusion in the Developing World." Washington, DC: World Bank, 2008.

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. Every ethical issue, at some level of generality, implicates some related legal issue. For instance, the classic "trolley problem" from philosophy has translated into the very urgent need to decide what is legally defensible when an autonomous vehicle is faced with an accident that might harm human beings. Certain decisions which would be acceptable for a human being would not necessarily be tolerated by society when taken by AI or embedded in AIs. In this sense, the recommendations of the Law Committee should be understood as an important complement to the ethics recommendations provided by other Committees. Additionally, we are concerned that some humans are particularly vulnerable in this area, for example children and those with mental and physical disabilities.

The development, design, and distribution of AI/AS should fully comply with all applicable international and domestic law. This obvious and deceptively simple observation obscures the many deep challenges AI/AS pose to legal systems; global-, national-, and local-level regulatory capacities; and individual rights and freedoms.

Our concerns and recommendations fall into three principal areas:

1. Governance and liability
2. Societal impact
3. "Human in the loop"

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, and domestic and international legislation in these areas and we welcome this opportunity given to us by The IEEE Global Initiative to ensure that the huge benefits available to humanity and our planet from AI/AS are thoughtfully stewarded for the future.

8 Law

Issue:

How can we improve the accountability and verifiability in autonomous and intelligent systems?

Background

Most users of AI systems will not be aware of the sources, scale, and significance of uncertainty in AI systems' operations. The proliferation of AI/AS will see an increase in the number of systems that rely on machine learning and other developmental systems whose actions are not pre-programmed and that do not produce "logs" of how the system reached its current state. This process creates difficulties for everyone ranging from the engineer to the lawyer in court, not to mention ethical issues of ultimate accountability.

Candidate Recommendations

Although we acknowledge this cannot be done currently, AI systems should be designed so that they always are able, when asked, to show the registered process which led to their actions to their human user, identify any sources of uncertainty, and state any assumptions they relied upon.

Although we acknowledge this cannot be done currently, AI systems should be programmed so that they proactively inform users of such uncertainty even when not asked under certain circumstances.

With higher potential risk of economic or physical harm, there should be a lower threshold for proactively informing users of risks and a greater scope of proactive disclosure to the user.

Designers should leverage current computer science regarding accountability and verifiability for code.

Lawmakers on national, and in particular on international, levels should be encouraged to consider and carefully review a potential need to introduce new regulation where appropriate, including rules subjecting the market launch of new AI/AS driven technology to prior testing and approval by appropriate national and/or international agencies.

Further Resources

1. Kroll, Joshua. "Accountable Algorithms." PhD diss., Princeton, NJ: Princeton University, 2015.

2. Datta, Anupam, Shayak Sen, and Yair Zick. "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems." 2016 IEEE Symposium on Security and Privacy, May 22–26, 2016.DOI: 10.1109/SP.2016.42.

8 Law

Issue:

How to 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.

Background

Government increasingly automates part or all of its decision-making. Law mandates transparency, participation, and accuracy in government decision-making. When government deprives individuals of fundamental rights individuals are owed notice and a chance to be heard to contest those decisions. A key concern is how legal commitments of transparency, participation, and accuracy can be guaranteed when algorithmic-based AI systems make important decisions about individuals.

Candidate Recommendations

1. Governments should not employ AI/AS that cannot provide an account of the law and facts essential to decisions or risk scores. The determination of, for example, fraud by a citizen should not be done by statistical analysis alone. Common sense in the AI/AS and an ability to explain its logical reasoning must be required. All decisions taken by governments and any other state authority should be subject to review by a court, irrespective of whether decisions involve the use of AI/AS technology. Given the current abilities of AI/AS, under no circumstances should court decisions be made by such systems. Parties, their lawyers, and courts must have access to all data and information generated and used by AI/AS technologies employed by governments and other state authorities.
2. AI systems should be designed with transparency and accountability as primary objectives. The logic and rules embedded in the system must be available to overseers of systems, if possible. If, however, the system's logic or algorithm cannot be made available for inspection, then alternative ways must be available to uphold the values of transparency. Such systems should be subject to risk assessments and rigorous testing.
3. Individuals should be provided a forum to make a case for extenuating circumstances that the AI system may not appreciate—in other words, a recourse to a human appeal. Policy should not be automated if it has not undergone formal or informal rulemaking procedures, such as interpretative rules and policy statements.
4. Automated systems should generate audit trails recording the facts and law supporting decisions. Audit trails should include a

8 Law

comprehensive history of decisions made in a case, including the identity of individuals who recorded the facts and their assessment of those facts. Audit trails should detail the rules applied in every mini-decision made by the system.

Further Resources

- Schwartz, Paul. "Data Processing and Government Administration: The Failure of the American Legal Response to the Computer." *Hastings Law Journal* 43 (1991): 1321–1389.
- Citron, Danielle Keats. "Technological Due Process." *Washington University Law Review* 85 (2007): 1249–1313.
- Citron, Danielle Keats. "Open Code Governance." *University of Chicago Legal Forum* (2008): 355.
- Crawford, Kate, and Jason Schultz. "Big Data and Due Process: Toward a Framework to Address Predictive Privacy Harms." *Boston College Law Review* 55 (2014): 93.
- Pasquale, Frank. *Black Box Society*. Harvard University Press, 2014.
- Bamberger, Kenneth. "Technologies of Compliance: Risk and Regulation in the Digital Age." *Texas Law Review* 88 (2010): 699.
- Kroll, Joshua. *Accountable Algorithms*. Princeton, NJ: Princeton University Press, 2015.

Issue:

How can AI systems be designed to guarantee legal accountability for harms caused by these systems?

Background

One of the fundamental assumptions most laws and regulations rely on is that human beings are the ultimate decision makers. As autonomous devices and AI become more sophisticated and ubiquitous, that will increasingly be less true. The AI industry legal counsel should work with legal experts to identify the regulations and laws that will not function properly when the "decision-maker" is a machine and not a person.

Candidate Recommendations

Any or all of the following can be chosen. The intent here is to provide as many options as possible for a way forward for this principle.

1. Designers should consider adopting an identity tag standard—that is, no agent should be released without an identity tag to maintain a clear line of legal accountability.
2. Lawmakers and enforcers need to ensure that the implementation of AI systems is not abused as a means to avoid liability of those businesses and entities employing the AI. Regulation should be considered to require a sufficient capitalization or insurance guarantee of an AI system that could be held liable for injuries and damages caused by it.

8 Law

3. In order to avoid costly lawsuits and very high standards of proof that may unreasonably prevent victims from recovering for damages caused by AI, states should consider implementing a payment system for liable AI similar to the worker's compensation system. The standard of evidence necessary to be shown to recover from the payment system would be lower: victims only need to show actual injury or loss and reasonable proof that the AI caused the injury or loss. But in return for easier and faster payments, the payments would be lower than what might be possible in court. This permits the victims to recover faster and easier while also letting AI developers and manufacturers plan for an established potential loss.

4. Companies that use and manufacture AI should be required to establish written policies governing how the AI should be used, who is qualified to use it, what training is required for operators, and what operators and other people can expect from the AI. This will help to give the human operators and beneficiaries an accurate idea of what to expect from the AI while also protecting the companies that make the AI from future litigation.

5. States should not automatically assign liability to the person who turns on the AI. If it is appropriate to assign liability to a person involved in the AI's operation, it is most likely the person who oversees or manages the AI while it operates, who is not necessarily the person who turned it on.

6. Human oversight of AI should only be required when the primary purpose of the AI is to improve human performance or eliminate human error. When the primary purpose of the AI is to provide for human convenience, like autonomous cars, requiring oversight defeats the purpose of the AI.

7. Intellectual property statutes should be reviewed to clarify whether amendments are required in relation to the protection of works created by the use of AI. The basic rule should be that when an AI product relies on human interaction to create new content or inventions, the human user is the author or inventor and receives the same intellectual property protection as if he or she had created the content or inventions without any help from AI.

Further Resources

- Weaver, John Frank. Robots Are People Too: How Siri, Google Car, and *Artificial Intelligence Will Force Us to Change Our Laws*. Praeger, 2013.

8 Law

Issue:

How can autonomous and intelligent systems be designed and deployed in a manner that respects the integrity of personal data?

Background

AI heightens the risk regarding the integrity of personal data. As consumers, we are worried about privacy but also about the integrity of our data, including the danger of our data being hacked, misused, or even falsified. This is not a concern that is unique to AI, but AI heightens it.

Candidate Recommendation

1. Generally, encourage research/measures/products aiming to ensure data integrity; clarify who owns which data in which situations.
2. Discuss regulation and the pros and cons of regulation of data ownership by individuals and companies.

Further Resources

- Pasquale, Frank. *Black Box Society*. Harvard University Press, 2014.
- Artificial Intelligence, Robotics, Privacy, and Data Protection, 38th International Conference of Data Protection and Privacy Commissioners, 2016.

New Committee Descriptions

Classical Ethics in Information and Communication Technologies

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

The following are the Subcommittees of the Classical Ethics in Information Communications Technology (ICT) Committee along with sample Issues being created for the next version of *Ethically Aligned Design*.

- ### Function, purpose, identity, and agency.

  Issue: How can classical ethics act as a regulating force in autonomous technologies as goal-directed behavior transitions from being exogenously set—one that is given by operators at set points in the development and implementation cycle—to one that is indigenously set in situ? A virtue ethics approach has merits for accomplishing this even without having to posit a "character" in an autonomous technology, since it places emphasis on habitual, iterative action focused on achieving excellence in a chosen domain or in accord with a guiding purpose.

- ### Creation of an Agenda for Ethics in Start-ups and Tech Giants.

  Issue: Though ethical considerations in tech design are considered "increasingly important" by companies, they are "not quite on the agenda yet." How can ethical considerations be prioritized among start-up and tech giant companies, public projects, and research consortiums? What place does classical ethics have in such an agenda?

- ### From Committee to Action

  Issue: Classical ethics can go a long way toward exploring concerns, but do not always offer concrete solutions for researchers, innovators, and business. Is classical ethics accessible and applicable in regards to tech projects? Perhaps the traditional classical ethical theories might not be adequate for the task at hand—that of informing the value design of machines. A meta-analysis of "classical ethics" is needed to address posited goals.

  The concept of responsible research and innovation (RRI) is a growing area, particularly within the EU, and is being adopted by research funders such as the EPSRC, who include the core principles in their mission statement. RRI is an umbrella concept that draws on classical ethics theory to provide tools and an approach that applies these principles to address ethical concerns from the outset of a project (design stage and onward).

New Committee Descriptions

- **Responsibility and Accountability via Classical Ethics Methodologies**

  Issue: How can classical ethics speak to issues of responsibility and accountability for autonomous applications whose abstracted layers of functionality are currently beyond our skill to collectively understand—thousands and thousands of lines of code, the working of which no one can verify, many people responsible for different parts, etc.—and for software which is supposed to learn and modify its own workings, and eventually modify itself.

- **Addressing Cultural Bias in the Design of Autonomous Systems (AS).**

  Issue: What are the fundamental values imposed on a system; what/which set(s) of values guide the design, and whether—if without consideration of non-Western values—artificial intelligence and autonomous systems (AI/AS) will generate problematic (e.g., discriminatory) consequences. There is an urgent need to broaden "traditional" ethics beyond the scope of "Western" ethics, e.g., utilitarianism, deontology, and virtue ethics; and include other traditions of ethics, e.g., Buddhism, Confucianism, etc.

  Data protection and privacy implications
  Issue: Human-produced details are included in big data, then input into AS, where the systems subsequently make decisions on our behalf and target us directly. Human decision-making is a complex and nuanced activity that AI is (as yet) unable to emulate. This makes the use of autonomous systems of concern due to the opportunity for error and negative outcomes. Does the centuries-old tradition of classical ethics have something unique to offer toward understanding the human decision-making process that could then inform data protection and privacy issues?

- **Anthropomorphistic approaches toward Internet technology (IT), ICTs, and AI**

  Issue: The current approach to autonomous systems often erroneously blurs the distinction between agents and patients (Capurro), interpreted as a distinction between "natural" self-organizing systems and artificial, non-self-organizing devices. The attempt to implant human morality and human emotion into AI is a misguided attempt to designing value-based systems.

- **Ethics vocabulary (and subsequent entrenchment of values)**

  Issue: Philosophers and ethicists are trained in vocabulary relating to philosophical concepts and terminology. There is an intrinsic value placed on these concepts since it is seen as fundamental and foundational to the concept. Using this vocabulary in "real life" instances does not function as well since not everyone has been trained to comprehend the nuances. However, not understanding a philosophical definition does not detract from the necessity of its being. How can classical ethics address this tension?

New Committee Descriptions

Mixed Reality

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.

The following are Issues the Mixed Reality Committee is focusing on for the next version of *Ethically Aligned Design*.

1. **Issue:** Within the coming realm of AI/AS-enhanced mixed reality, how can we evolve, harness, and not eradicate the positive effects of serendipity?

   Background: In the real world, bumping into a stranger when your GPS breaks means you may meet your life partner. However, in the digital and virtual spheres, algorithms that have been programmed by design may eliminate genuine randomness from our human experience. What do we stand to lose when we code "frictions" out of our lives that may cause discomfort but also joy and growth?

2. **Issue:** What are the connections between the physical and the psychological, the body and mind? How can AI-enhanced mixed reality explore these connections for therapeutic and other purposes, and what are the risks?

   Background: Being in a completely mediated VR environment could, for example, fool the mind into thinking and feeling as it did in an earlier stage of one's life, with measurable physiological effects. Psychological conditions often have accompanying physical ailments that diminish or disappear when the psychological condition is treated. How can MR be used constructively to engage the mind to such an extent that physiological mechanisms can be controllably affected, and what are the ethical implications?

3. **Issue:** When an AI-based mixed-reality system controls the senses, does it control the mind? What are the short- and long-term effects and implications of giving over one's senses to software?

New Committee Descriptions

Background: A VR system can radically affect how the mind processes and synthesizes information, and ultimately it could be a way to teach ourselves new ways to think and create content. However, the long-term effects of immersion are largely unknown at this point, and the exploitability of a person's (or a larger group's) notion of reality raises a host of ethical issues.

4. **Issue:** What happens to cultural institutions in an AI-enabled world of illusion, where geography is largely eliminated, tribe-like entities and identities could spring up spontaneously, and the notion of identity morphs from physical certainty to virtuality?

   Background: When an increasing amount of our lives is spent in a photorealistic and responsive world of software, what will happen to actual human contact, which might always remain un-digitizable in meaningful ways? When an illusory world is vastly more pleasant and fulfilling than the physical alternative, will there be a significant population who choose to live in a synthetic world of their own making? Opting in and out will be central to the coming digital experiences; but what happens with the opposite—when people choose to opt-out of the "real" world in favor of illusion?

5. **Issue:** A mixed-reality world driven by intelligent systems would have to be observing behavior and sensing physical phenomena continuously in order to provide individuals with appropriate content. Is that a sensible trade-off or a dance with the devil?

   Background: Does the sensing possible with these technologies allow for ways to level the playing field (e.g., Steve Mann's notion of "sousveillance"), or does it exacerbate existing power inequities? If the AI-enabled world of the future is monetized with personal data, then surveillance might well be baked into the system, but the ethical issues we see today will only be amplified in an immersive environment.

New Committee Descriptions

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 well-being in line with innovation.

The following are the Subcommittees of The Affective Computing Committee along with sample Issues being created for the next versions of *Ethically Aligned Design*.

1. **Systems supporting human potential. Addresses human flourishing, dignity, human autonomy, needs satisfaction, nudging for social good.**

   Issue: We are concerned for a catastrophic loss of individual human autonomy. Some systems may negatively affect human psychological wellbeing.

2. **When systems lie. Addresses honest signals, trust, and deception. Transparency.**

   Issue: Should we, and if so how do we, regulate computing and robotic artifacts that are able to tap into the affective system of humans in terms of who benefits, vulnerable populations, and human rights, and to what extent do the policies, guidelines, and laws that already exist ensure that ethical behavior is adhered to by the designers and providers of these systems.

3. **When systems become intimate. Addresses intimacy and relations with machines.**

   Issue: Concern exists for any efforts to develop intimate robots that will contribute to gender inequalities in society. Also, concern with respect to potential therapeutic use of intimate robots, e.g., recidivism in sex offenders—can this technology assist?

New Committee Descriptions

4. **When systems go across cultures. Addresses respect for cultural nuances of signaling where the artifact must respect the values of the local culture.**

   Issue: Affective systems should not affect negatively the cultural/socio/religious values of the community where they are inserted. We should deploy affective systems with values that are not different from those of the society where they are inserted.

5. **When systems have their own "feelings." Addresses robot emotions, moral agency and patiency, and robot suffering.**

   Issue: Deliberately constructed emotions are designed to create empathy between humans and artifacts, which may be useful or even essential for human-AI collaboration. However, this could lead humans to falsely identify with the AI. Potential consequences are over-bonding, guilt, and above all: misplaced trust.

6. **System manipulation. Addresses when systems manipulate emotions to alter human behavior and use emotions to sell us stuff, subtly or overtly; also with respect to the question of transparency.**

   Issue: There is an open question whether system manipulation (or nudging) of people using affect is appropriate. Is it acceptable when the global community benefits exceed individual benefits? Or are there fundamental individual rights that transcend these utilitarian arguments?

New Committee Descriptions

Effective Policymaking for Innovative Communities Involving Artificial Intelligence and Autonomous Systems (EpicAI)

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 experimental to support the innovation of AI/AS for shared human benefit.

A cornerstone of the Committee's approach is bringing together policy with the practical considerations of industry. Doing so provides two benefits:

- For policymakers, strong connection with product developers ensures that policy can remain relevant and practical. Policy developed in isolation risks becoming impractical to implement or too slow to keep up with changes in the market and among technology users.
- For industry, the connection to policymakers ensures that narrow financial and competitive interests do not overwhelm the need for ethical behavior in regard to the development and application of these technologies.

The Committee believes that achieving realistic and fair guidelines requires cooperation between both policymakers and commercial interests. The IEEE Global Initiative can help foster a multi-sector perspective on artificial intelligence and autonomous systems (AI/AS).

The EPICAI committee is currently focusing on the following two Issues:

1. **Issue:** How can we help public service entities (governments, non-profits, public-private partnerships, and members of the public) more rapidly adopt AI/AS to improve public service?

   Background: Today, many members of public service may be unfamiliar with what

New Committee Descriptions

AI/AS can do. It is important to inform and educate representatives about AI/AS to best inform future innovative policymaking. There are significant opportunities to improve dramatically the quality of public service provided by using AI/AS for the rote, rule-based, routine functions of traditional government organizations. Incorporating AI/AS paired with human activities into such functions will help educate and train a new generation of public service professionals on the innovation possibilities, benefits, and limitations of these technologies. IEEE can help ensure these efforts are guided by appropriate ethical principles.

2. **Issue:** How can we best create recommendations to help the private and public sectors collaborate to explore innovative uses of AI/AS, without excessive limitations, for the betterment of society?

Background: Completing this objective will present a balanced perspective that recognizes both sectors are co-dependent, requiring multi-sector partnerships to prevent the misuse of AI/AS while ensuring that innovation improves the lives of all individuals. It is not sufficient for AI/AS to only improve select organizations or some people, and our work is designed to help ensure these efforts are guided by appropriate ethical principles. A particular focus will be placed on the case for combined human and AI/AS pairings, optimizing the strengths of both humans and AI/AS together (versus an either-or proposition), and minimizing the weaknesses.

End Notes

General Principles

[i] UN General Assembly. "Universal Declaration of Human Rights." Paris. 1948. Available at http://www.un.org/en/universal-declaration-human-rights/.

[ii] UN General Assembly. 1966. "International Covenant on Civil and Political Rights." 1966. Treaty Series 999 (December): 171. Available at http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx.

[iii] UN General Assembly. 1989. "Convention on the Rights of the Child." Treaty Series 1577 (November): 3. Available at http://www.ohchr.org/en/professionalinterest/pages/crc.aspx.

[iv] UN General Assembly. 1979. "Convention on the Elimination of All Forms of Discrimination against Women." Treaty Series 1249: 13. Available at http://www.un.org/womenwatch/daw/cedaw/.

[v] UN General Assembly. 2006. "Convention on the Rights of Persons with Disabilities." Treaty Series 2515 (December): 3. Available at http://www.un.org/disabilities/convention/conventionfull.shtml.

[vi] Council on Foreign Relations. "Geneva Conventions." Accessed November 10, 2016. http://www.cfr.org/human-rights/geneva-conventions/p8778.

Embedding Values into Autonomous Intelligent Systems

[vii] Velasquez, Manuel and Claire Andre, Thomas Shanks, S.J., and Michael J. Meyer. "The Common Good." Issues in Ethics, 5, no. 1 (Spring 1992). Available at http://www.scu.edu/ethics/publications/iie/v5n1/common.html.

[viii] Kahn, Peter H., Jr., Aimee L. Reichert, Heather E. Gary, Takayuki Kanda, Hiroshi Ishiguro, Solace Shen, Jolina H. Ruckert, and Brian Gill. "The New Ontological Category Hypothesis in Human-Robot Interaction." In Proceedings of the 6th International Conference on Human-Robot Interaction, 159–160. HRI '11. New York, NY, USA: ACM, 2011. doi:10.1145/1957656.1957710. Available at http://dl.acm.org/citation.cfm?id=1957656.1957710.

[ix] Robinette, Paul, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. "Overtrust of Robots in Emergency Evacuation Scenarios." In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 101–108. IEEE, 2016. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7451740.

[x] Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." arXiv Preprint arXiv:1607.06520, 2016. Available at https://arxiv.org/abs/1607.06520.

End Notes

[xi] University of Washington, DO-IT. "DO-IT | Disabilities, Opportunities, Internetworking, and Technology." Accessed November 10, 2016. http://www.washington.edu/doit/.

[xii] Wallach, Wendell, and Colin Allen. Moral Machines: Teaching Robots Right from Wrong. 1 edition. New York; Oxford: Oxford University Press, 2010. Available at https://www.amazon.com/Moral-Machines-Teaching-Robots-Right/dp/0199737975.

[xiii] Weng, Yueh-Hsuan, Yusuke Sugahara, Kenji Hashimoto, and Atsuo Takanishi. "Intersection of 'Tokku' Special Zone, Robots, and the Law: A Case Study on Legal Impacts to Humanoid Robots." International Journal of Social Robotics 7, no. 5 (November 1, 2015): 841–57. doi:10.1007/s12369-015-0287-x.

[xiv] Regulation 2016/679 of the European Parliament and of the Council on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Directive), 2016 O.J. L 119/1. Available at http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf.

[xv] Winfield, Alan. "On Internal Models, Consequence Engines and Popperian Creatures." Robohub, September 8, 2014. http://robohub.org/on-internal-models-consequence-engines-and-popperian-creatures/.

Methodologies to Guide Ethical Research and Design

[xvi] Capurro, Rafael. "Intercultural Information Ethics." In Case Studies in Library and Information Science Ethics, edited by Elizabeth A. Buchanan, 11:10. Mcfarland & Co., 2008. http://www.capurro.de/iie.html.

[xvii] Ewel, Jim. "What Is Agile Marketing?" Agile Marketing. Accessed November 10, 2016. http://www.agilemarketing.net/what-is-agile-marketing/.

[xviii] CSER Cambridge. Kay Firth-Butterfield: Lucid AI's Ethics Advisory Panel., 2016. https://www.youtube.com/watch?v=w3-wYGbNZU4.

[xix] "Code of Conduct for BCS Members." BCS - The Chartered Institute for IT, June 3, 2015. http://www.bcs.org/category/6030.

[xx] IBM MobileFirst. Watson Demo: Doctor's Consultation, 2014. https://www.youtube.com/watch?v=lRhg6yxenY4.

[xxi] "Home Page for Professor Michael Kearns, University of Pennsylvania." Last modified November 10, 2016. https://www.cis.upenn.edu/~mkearns/#publications.

[xxii] Committee on Legal Affairs, European Parliament. Draft Report on Civil Law Rules on Robotics 2015/2103(INL). May 2016. Available at http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN.

End Notes

Safety And Beneficence Of AGI & ASI

[xxiii] Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies. 1 edition. Oxford: Oxford University Press, 2014.

[xxiv] Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete Problems in AI Safety." arXiv Preprint arXiv:1606.06565, 2016. https://arxiv.org/abs/1606.06565.

[xxv] Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies.

[xxvi] Yudkowsky, Eliezer. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In Global Catastrophic Risks, edited by Nick Bostrom and Milan M. ĆirkoviĆ, 308–345. New York: Oxford University Press 2008.

[xxvii] Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." Minds and Machines 22, no. 2 (2012): 71–85. doi:10.1007/s11023-012-9281-3. http://link.springer.com/article/10.1007/s11023-012-9281-3.

[xxviii] Omohundro, Stephen M. "The Basic AI Drives." In Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference, 483–492. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2008. http://dl.acm.org/citation.cfm?id=1566174.1566226.

[xxix] Schneier, Bruce. "The Security Mindset." Schneier on Security, March 25, 2008. https://www.schneier.com/blog/archives/2008/03/the_security_mi_1.html.

[xxx] Christiano, Paul. "Security and AI Control." AI Control, October 15, 2016. https://medium.com/ai-control/security-and-ai-control-675ace05ce31#.90y7zmwec.

[xxxi] Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete Problems in AI Safety." arXiv Preprint arXiv:1606.06565, 2016. https://arxiv.org/abs/1606.06565.

[xxxii] Taylor, Jessica, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. "Alignment for Advanced Machine Learning Systems." Machine Intelligence Research Institute, 2016. https://intelligence.org/files/AlignmentMachineLearning.pdf.

[xxxiii] Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative Inverse Reinforcement Learning. Submission, 2016. https://arxiv.org/abs/1606.03137.

[xxxiv] Babcock, James, János Kramár, and Roman Yampolskiy. "The AGI Containment Problem." In Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings, edited by Bas Steunebrink, Pei Wang, and Ben Goertzel, 53–63. Cham: Springer International Publishing, 2016. http://dx.doi.org/10.1007/978-3-319-41649-6_6.

[xxxv] Yampolskiy, Roman. "Leakproofing the Singularity Artificial Intelligence Confinement Problem." Journal of Consciousness Studies 19, no. 1–2 (January 1, 2012): 194–214.

[xxxvi] Siddiqui, Md Amran, Alan Fern, Thomas G. Dietterich, and Shubhomoy Das. "Finite

End Notes

Sample Complexity of Rare Pattern Anomaly Detection." In Uncertainty in Artificial Intelligence: Proceedings of the 32nd Conference (UAI-2016), Edited by Alexander Ihler and Dominik Janzing, 686–695, 2016. https://pdfs.semanticscholar.org/aa25/53fdd6f9f48ea0ffb507edba4bdcb3e11470.pdf.

[xxxvii] Soares, Nate, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. "Corrigibility." In Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015. http://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10124.

[xxxviii] Armstrong, M. S., and L. Orseau. "Safely Interruptible Agents." ORA Review Team, May 2016. https://ora.ox.ac.uk/objects/uuid:17c0e095-4e13-47fc-bace-64ec46134a3f.

[xl] "Technical Debt." Wikipedia, September 28, 2016. https://en.wikipedia.org/w/index.php?title=Technical_debt&oldid=741642786.

[xli] "Null-Terminated String." Wikipedia, November 3, 2016. https://en.wikipedia.org/w/index.php?title=Null-terminated_string&oldid=747645477.

[xlii] "Buffer Overflow." Wikipedia, October 24, 2016. https://en.wikipedia.org/w/index.php?title=Buffer_overflow&oldid=746021516.

[xliii] Christiano, Paul. "Scalable AI Control." AI Control, December 5, 2015. https://medium.com/ai-control/scalable-ai-control-7db2436feee7#.agtb5klk9.

[xliv] Sandberg, Anders. "Ethics of Brain Emulations." Journal of Experimental & Theoretical Artificial Intelligence 26, no. 3 (July 3, 2014): 439–57. doi:10.1080/0952813X.2014.895113. http://www.tandfonline.com/doi/abs/10.1080/0952813X.2014.895113.

[xlv] Yampolskiy, Roman, and Joshua Fox. "Safety Engineering for Artificial General Intelligence." Topoi 32, no. 2 (2013): 217–26. doi:10.1007/s11245-012-9128-9. http://link.springer.com/article/10.1007%2Fs11245-012-9128-9.

[xlvi] Havens, John C. "Creating a Code of Ethics for Artificial Intelligence." Mashable, October 3, 2015. http://mashable.com/2015/10/03/ethics-artificial-intelligence/.

[xlvii] Partnership on Artificial Intelligence to Benefit People and Society. Accessed November 10, 2016. https://www.partnershiponai.org/.

[xlviii] Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies.

[xlix] Ibid.

[l] Yudkowsky, Eliezer. "Artificial Intelligence as a Positive and Negative Factor in Global Risk."

[li] Bostrom, Nick. "Strategic Implications of Openness in AI Development." Working paper, 2016. http://www.nickbostrom.com/papers/openness.pdf.

[lii] Partnership on Artificial Intelligence to Benefit People and Society. Accessed November 10, 2016. https://www.partnershiponai.org/.

[liii] Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies.

End Notes

Personal Data and Individual Access Control

[liv] Regulation 2016/679 of the European Parliament and of the Council on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Directive), 2016 O.J. L 119/1. Available at http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf.

[lv] Scoble, Robert. "I'm inside Tim Cook's Head but I Really Wonder What's Going on inside Mark Zuckerberg's Head?" Virtual Reality Pop, October 22, 2016. https://virtualrealitypop.com/im-inside-tim-cook-s-head-but-i-really-wonder-what-s-going-on-inside-mark-zuckerberg-s-head-5babf01c5713#.fe068u8du.

[lvi] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 1995 O.J. L 281, 31 – 50. Available at http://www.europa.eu.int/eur-lex/en/lif/dat/1995/en_395L0046.html.

[lvii] Lee, Phil. "Getting to Know the GDPR, Part 1 - You May Be Processing More Personal Information than You Think." Field Fisher Privacy and Information Law, October 2, 2015. http://privacylawblog.fieldfisher.com/2015/getting-to-know-the-gdpr-part-1-you-may-be-processing-more-personal-information-than-you-think/.

[lviii] United States v. Jones, 132 S. Ct 945 (2012). Available at https://www.supremecourt.gov/opinions/11pdf/10-1259.pdf.

[lix] General Data Protection Directive, 2016 O.J. L 119/1.

[lx] Greenberg, Andy. "Apple's 'Differential Privacy' Is About Collecting Your Data—But Not Your Data." Wired, June 13, 2016. https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/.

[lxi] Madden, Mary and Lee Rainie. Pew Research Center. "Americans' Attitudes About Privacy, Security and Surveillance." May 20, 2015. Available at http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/.

Reframing Autonomous Weapons Systems

[lxii] U.S. Department of Defense. "Directive on Autonomy in Weapon Systems, DoDD 3000.09." November 21, 2012. Available at http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.

[lxiii] "2010 Flash Crash." Wikipedia, October 26, 2016. https://en.wikipedia.org/w/index.php?title=2010_Flash_Crash&oldid=746360910.

Economics/Humanitarian Issues

[lxiv] GSMA. Half the World's Population Connected to the Mobile Internet by 2020, According to New GSMA Figures, November 6, 2014. http://www.gsma.com/newsroom/press-release/half-worlds-population-connected-mobile-internet-2020-according-gsma/.

Law

(No end notes included)

Global Initiative Membership

Executive Committee Members

(As of 13 December 2016)

Raja Chatila, Executive Committee Chair, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

Raja Chatila, IEEE Fellow, is Director of Research at the French National Center of Scientific Research (CNRS), and Director of the Institute of Intelligent Systems and Robotics (ISIR) at Pierre and Marie Curie University in Paris, France. He is also Director of the Laboratory of Excellence "SMART" on human-machine interactions.

His work covers several aspects of autonomous and interactive Robotics, in robot navigation and SLAM, motion planning and control, cognitive and control architectures, human-robot interaction, and robot learning, and to applications in the areas of service, field and space robotics. He is author of over 150 international publications on these topics. He is past President of the IEEE Robotics and Automation Society (2014-2015) and member of the French Commission on the Ethics of Research in Digital Sciences and Technologies (CERNA). IEEE Pioneer Award in Robotics and Automation.

Kay Firth-Butterfield, Executive Committee Vice-Chair, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

Kay Firth-Butterfield is a Barrister-at-Law and part-time Judge in the United Kingdom where she has also worked as a mediator, arbitrator, business owner and Professor of Law. In the United States, Kay is Executive Director and founding advocate of AI Austin and an adjunct Professor of Law. She is a humanitarian with a strong sense of social justice and has advanced degrees in Law and International Relations. Kay advises governments, think tanks, businesses and nonprofit organizations about artificial intelligence, law and policy.

Kay co-founded the Consortium for Law and Policy of Artificial Intelligence and Robotics at the University of Texas and teaches: Artificial Intelligence and emerging technologies: Law and Policy. She thinks about how AI and other emerging technologies will impact business and society, including how business can prepare for that impact in its internal planning and external interaction with customers and other stakeholders and how society will be affected by, and should prepare for, these technologies. She has a particular interest in AI and foreign policy and the use of AI to do good globally. Kay speaks regularly to international audiences addressing many aspects of these challenging changes.

Global Initiative Membership

John C. Havens, Executive Director of The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

John C. Havens is a regular contributor on issues of technology and wellbeing to Mashable, The Guardian, HuffPo and TechCrunch and is author of Heartificial Intelligence: Embracing Our Humanity To Maximize Machines and Hacking Happiness: Why Your Personal Data Counts and How Tracking it Can Change the World.

John was an EVP of a Top Ten PR Firm, a VP of a tech startup, and an independent consultant where he has worked with clients such as Gillette, P&G, HP, Wal-Mart, Ford, Allstate, Monster, Gallo Wines, and Merck. He is also the Founder of The Happathon Project, a non-profit utilizing emerging technology and positive psychology to increase human wellbeing.

John has spoken at TEDx, at SXSW Interactive (six times), and as a global keynote speaker for clients like Cisco, Gillette, IEEE, and NXP Semiconductors. John was also a professional actor on Broadway, TV and Film for fifteen years.

Dr. Greg Adamson, President, IEEE Society on Social Implications of Technology

Dr. Greg Adamson is President of the IEEE Society on Social Implications of Technology (SSIT) 2015-16. SSIT has addressed issues of ethics and emerging technologies throughout its 44-year history. Greg also chairs the Ethics, Society and Technology Committee of the IEEE Technical Activities Board, which oversees the IEEE TechEthics™ program. He is an Associate Professor at the University of Melbourne School of Engineering, and a consultant in cybercrime and blockchain with Digital Risk Innovation. He initiated the IEEE Conference on Norbert Wiener in the 21st Century series, (Boston 2014, Melbourne 2016, Budapest 2018). Wiener's 20th century work in both the technical and social impact of technology fields foreshadowed current discussions in ethics in the design of autonomous systems. Greg was the invited keynote on Ethics at the World Engineering Conference and Convention in Kyoto November 2015. His primary area of research interest is barriers to the uptake of socially beneficial technology.

Ronald C. Arkin, Regents' Professor, Associate Dean for Research in the College of Computing at Georgia Tech

Ronald C. Arkin is Regents' Professor and Associate Dean for Research in the College of Computing at Georgia Tech and is the Director of the Mobile Robot Laboratory. He served as STINT visiting Professor at KTH in Stockholm, Sabbatical Chair at the Sony IDL in Tokyo, and in the Robotics and AI Group at LAAS/CNRS in Toulouse. Dr. Arkin's research interests include behavior-based control and action-oriented perception for mobile robots and UAVs, deliberative/reactive architectures, robot survivability, multiagent robotics, biorobotics, human-robot interaction, machine deception, robot ethics, and learning in autonomous systems. His books include Behavior-Based Robotics (MIT Press), Robot Colonies (Kluwer), and Governing Lethal Behavior in Autonomous Robots (Taylor & Francis). He has provided expert testimony to the United Nations, the International Committee of the Red Cross, the Pentagon and

Global Initiative Membership

others on Autonomous Systems Technology. Prof. Arkin served on the Board of Governors of the IEEE Society on Social Implications of Technology, the IEEE Robotics and Automation Society (RAS) AdCom, and is a founding co-chair of IEEE RAS Technical Committee on Robot Ethics. He is a Distinguished Lecturer for the IEEE Society on Social Implications of Technology and a Fellow of the IEEE.

Stephen L. Diamond, Global Standards Officer and General Manager, Industry Standards Office, EMC Corporation

Stephen L. Diamond is Global Standards Officer and General Manager of the Industry Standards Office at EMC Corporation, where he is responsible for industry standards engagements and related intellectual property. Prior to EMC, Steve was Director of Product Management for Intercloud Computing at Cisco Systems. Before that, he was Vice President of Marketing at Equator Technologies, a DSP startup. His professional interests include IoT, Big Data, cyberphysical systems, brain/machine interface, AI, Intercloud computing, the "third platform," and related industry standards developments.

Presently, Steve is Vice-chair of the IEEE Future Directions Committee, Chair of the IEEE Cloud Computing Standards Committee, and President Emeritus of the IEEE Computer Society. He was the Founder and Chair of the IEEE Cloud Computing Initiative from 2011-14, and served on the IEEE Board of Directors from 2005-06 and 2009-10. Steve was the 2003 President of the IEEE Computer Society and Editor-in-Chief of IEEE Micro Magazine from 1995-98.

Steve was awarded the IEEE Computer Society Richard E. Merwin Medal in 2014, the IEEE Third Millennium Medal in 2000, and the IEEE Computer Society Golden Core Award in 1997. For further details, visit https://www.linkedin.com/in/stephenldiamond or contact him at s.diamond (at) ieee.org.

Virginia Dignum, Associate Professor, Faculty of Technology Policy and Management, TU Delft

Virginia Dignum (female, 1964) is an Associate Professor at the Faculty of Technology Policy and Management at TU Delft. She holds a PhD from the Utrecht University, in 2014. Previously, she worked for more than 12 years in consultancy and system development in the areas of expert systems and knowledge management. Her research focuses on value-sensitive design of intelligent systems and multi-agent organizations, focusing on the formalization of moral and normative behaviors and social interactions. In 2006, she was awarded the prestigious Veni grant from NWO (Dutch Organization for Scientific Research) for her work on agent-based organizational frameworks. She has participated and reviewed several EU and national projects, is member of the reviewing boards for the main journals and conferences in AI and has chaired many international conferences and workshops. She has published more than 180 peer-reviewed papers and several books, yielding an h-index of 27. She is secretary of the International Foundation for Autonomous Agents and Multi-agent Systems (IFAAMAS) and co-chair of the European Conference on Artificial Intelligence (ECAI) in 2016.

Global Initiative Membership

Philip Hall, Member, IEEE-USA Government Relations Council

Philip Hall serves on the IEEE-USA Government Relations Council (GRC) as Chair of the IEEE-USA Committee on Transport and Aerospace Policy, and also as Co-Chair of the recently established IEEE-USA Ad Hoc Committee on Artificial Intelligence Policy (which also covers autonomous systems). He is also Vice President of the IEEE Society on Social Implications of Technology (SSIT), a member of the Working Group for IEEE P7000™ Model Process for Addressing Ethical Concerns During System Design and was General Co-Chair of the 2016 IEEE International Symposium on Ethics in Engineering, Science, and Technology (IEEE ETHICS 2016) held in Vancouver BC in May 2016.

In addition, Philip is a Principal Fellow ('Professor-at-Large') in the Department of Electrical and Electronic Engineering at The University of Melbourne (Australia), and co-Founder & CEO of RelmaTech Limited (UK), a company developing innovative technology solutions for the spatial management of semi-autonomous and autonomous vehicles. In 2013 he co-chaired a high-level roundtable sponsored by the Australian Government to consider the role of new technologies in Securing Australia's Future, and participated in the 2nd World Emerging Industries Summit in China as a guest of the Asia-Pacific CEO Association and the Wuhan Municipal Government. More recently he organized and chaired exclusive seminars in Melbourne, Australia and Washington DC to consider the National Security and Societal Implications of Remotely Piloted Airborne Vehicles and Related Technologies. He was a finalist judge for the United Arab Emirates 2015 and 2016 'Drones for Good' Awards and the 2016 'Robotics & AI for Good' Award, and is currently an associate editor for the international Journal of Unmanned Vehicle Systems.

Philip is also a member of the IEEE Aerospace & Electronics Systems Society (AESS) where he is Vice Chair of the AESS Unmanned Airborne Vehicle Technical Panel, a member of the Working Group for IEEE P1920.1™ Aerial Communications and Networking Standards, and Chair IEEE Region 4 Southeastern Michigan Chapter 3: Aerospace & Electronics Systems, Communications.

Eva Schulz-Kamm, Political Affairs & Public Co-Creation, NXP

Eva Schulz-Kamm heads NXP's Political Affairs & Public Co-Creation Team. She brings to the function a truly unique background. In addition to extensive work on research funding initiatives, marketing strategies, university relations, intellectual property proposals, and governmental standardization policies, she holds a degree in Physics, with a focus in solid-state physics and thermodynamics, from the University of Karlsruhe and an MBA in Innovation Management and Entrepreneurship from the EDHEC Business School in Nice, France.

Additionally, Eva Schulz-Kamm has held key positions in numerous companies and organizations including the Fraunhofer Institute for Solar Energy Systems, MVV Energie AG, ZukunftsAgentur Brandenburg GmbH, and the Federal Ministry of Education and Research. Most recently, she was a Director at DIHK e.V. Association of German Chambers of Industry

Global Initiative Membership

and Commerce where she was responsible for the development of national and international Innovation and Technology Policy including Standardization Policy.

Raj Madhavan, Founder & CEO, HumRobTech, LLC, USA & Distinguished Visiting Professor of Robotics, Amrita University, India

Raj Madhavan is the Founder & CEO of Humanitarian Robotics Technologies, LLC, Maryland, U.S.A. and a Distinguished Visiting Professor of Robotics with AMMACHI Labs at Amrita University, Kerala, India. He has held appointments with the Oak Ridge National Laboratory (March 2001-January 2010) as an R&D staff member based at the National Institute of Standards and Technology (March 2002-June 2013), and as an assistant and associate research scientist, and as a member of the Maryland Robotics Center with the University of Maryland, College Park (February 2010-December 2015). He received a Ph.D. in Field Robotics from the University of Sydney and an ME (Research) in Systems Engineering from the Australian National University.

Over the last 20 years, he has contributed to topics in field robotics, and systems and control theory. His current research interests lie in humanitarian robotics and automation – the application and tailoring of existing and emerging robotics and automation technologies for the benefit of humanity in a variety of domains, including unmanned (aerial, ground) vehicles in disaster scenarios. Dr. Madhavan has edited three books and four journal special issues, and has published over 185 papers in archival journals, conferences, and magazines. Within the IEEE Robotics and Automation Society, he served as the Founding Chair of the Technical Committee on Performance Evaluation and Benchmarking of Robotics and Automation Systems, TC-PEBRAS (2009-2011), Founding Chair of the Humanitarian Robotics and Automation Technology Challenge, HRATC (2014, 2015), Vice President of the Industrial Activities Board (2012-2015), Chair of the Standing Committee for Standards Activities (2010-2015), and since 2012 is the Founding Chair of the Special Interest Group on Humanitarian Technology (RAS-SIGHT). He is the 2016 recipient of the IEEE Robotics and Automation Society's Distinguished Service Award for his "distinguished service and contributions to RAS industrial and humanitarian activities".

Richard Mallah, Director of Artificial Intelligence Projects Future of Life Institute

Richard is Director of AI Projects at NGO the Future of Life Institute, where he works to support the robust, safe, beneficent development of both short-term and long-term artificial intelligence via analysis, metaresearch, organization, research direction, and advocacy. Issues tackled range from AI-induced unemployment and legal issues around autonomy to meaningful control and deep value alignment. Mallah also heads research in AI at enterprise knowledge integration platform

Global Initiative Membership

firm Cambridge Semantics, Inc., as Director of Advanced Analytics, leading R&D of technologies for knowledge representation, machine learning including deep learning, computational linguistics, conceptual middleware, and automated systems generation, powering applications ranging from fraud analytics to drug discovery.

He is an advisor to other nonprofits and companies where he advises on AI, knowledge management, and sustainability. He has over fifteen years' experience in AI algorithms development, product team management, and CTO-level roles. Richard holds a degree in intelligent systems and computer science from Columbia University.

AJung Moon, Co-founder Open Roboethics initiative

AJung Moon is a co-founder of the Open Roboethics initiative (ORi), an international roboethics think tank that investigates ways in which stakeholders of robotics technologies can work together to influence how robots should shape our future. What should a robot do? What decisions are we comfortable delegating to robots? These are some of the questions ORi has been exploring in the domain of self-driving vehicles, care robots, as well as lethal autonomous weapons systems.

AJung Moon is also a Vanier Scholar and a Ph. D. Candidate in Mechanical Engineering at the University of British Columbia. She designs nonverbal behaviors (hand gestures, gaze cues) that robots can use to better collaborate with people. As a roboticist, she believes that answering the question "What should a robot do?" can be a highly creative process that involve active discussions between designers and members of other stakeholder groups.

Monique Morrow, CTO New Frontiers Engineering at Cisco

Monique has a track record of co-innovating with customers that has transcended the globe from North America, Europe and Asia. Monique's current focus in on the intersection between research in economics and technology to portfolio execution (e.g. Circular and Exponential Economies) along with defining mechanisms and marketplace scenarios for cloud federation constructs to include security.

Monique is passionate about the humanitarian use of technology, in addition to exploring the use of AI/VR to create a people neutral system that is both experiential and ethical without losing the beauty of randomness in human behavior. She is focused on developing the use of blockchain to create identity as a service, applying humanistic and purposeful values in an organization, and creating modes of privacy that are understood by all members of our ecosystem.

Francesca Rossi, Full Professor of Computer Science University of Padova, Italy

Francesca Rossi is a full professor of computer science at the University of Padova, Italy, currently on leave at the IBM Research Center at Yorktown Heights, NY. Her research interests include constraint reasoning, preferences, multi-agent systems, computational social choice, and artificial intelligence. She has been president

Global Initiative Membership

of the international association for constraint programming (ACP) and of the international joint conference on AI (IJCAI). She has been program chair of severe international conferences, among which CP 2003 and of IJCAI 2013. She is a Radcliffe fellow, as well as an AAAI and an ECCAI fellow.

She is in the editorial board of Constraints, Artificial Intelligence, AMAI, and KAIS. She is the associate editor in chief of JAIR. She has published more than 170 papers and one book, and she co-edited 13 volumes, among which the handbook of constraint programming. She is co-chairing the AAAI committee on AI and ethics and she is in the scientific advisory board of the Future of Life Institute and of Insight (Ireland). She has given interviews on various topics related to AI to several media and newspapers, such as the Wall Street Journal, CNBS Italy, Motherboard, Science, the Washington Post, Sydney Morning Herald, Euronews, New Scientist, La Repubblica, Corriere della Sera. She has delivered three TEDx talks on Artificial Intelligence.

Dr. Yu Yuan, Founder, CATE Global

Dr. Yu Yuan is currently serving as the Chair of IEEE Digital Senses Initiative, the Chair of IEEE SCC42 Transportation (IEEE Standards Coordinating Committee on Transportation), the Chair of IEEE 2040 Working Group (Standard for Connected, Automated and Intelligent Vehicles), the Standards Chair of IEEE Consumer Electronics Society, and a Board Member and the SCC Coordinator of IEEE Standards Association Standards Board. He is a veteran researcher and practitioner in the areas of Transportation, Consumer Electronics, and Internet of Things. Dr. Yuan founded CATE Global, a multinational think tank focusing on bringing world-class expertise to clients and projects in China, and is serving as the President. He is also serving as the CEO of Motiveware Technology, a company developing and providing disruptive enabling technologies for mobile internet, location based services, connected vehicles and automated driving.

Prior to this he had been working for IBM Research and was also a key contributor to IBM IoT Technology Center. Dr. Yuan is experienced in the patent process and invention development. He has filed numerous patents and received many IBM Invention Achievement awards and IBM High Value Patent awards. Dr. Yuan has published extensively in referred conferences and journals, and he has organized many major conferences as general chair or program chair. Learn more at www.linkedin.com/in/DrYuYuan.

Global Initiative Membership

Committee Descriptions & Members
(As of 13 December 2016)

General Principles

This Committee is working on the overarching principles of the ethical design and use of autonomous and intelligent systems, focusing on the broader issues related to the acceptance of these systems by the general public.

- **Kay Firth-Butterfield** (Co-Chair) – Executive Director, AI Austin; Consortium on Law and Ethics of Artificial Intelligence and Robotics, Strauss Center, University of Texas; University of Texas Law School
- **Alan Winfield** (Co-Chair) – Professor, Bristol Robotics Laboratory, University of the West of England; Visiting Professor, University of York
- **Huw Price** – Bertrand Russell Chair of Philosophy, University of Cambridge; Centre for the Future of Intelligence, University of Cambridge
- **Seán Ó hÉigeartaigh** – Centre for the Study of Existential Risk, University of Cambridge; Centre for the Future of Intelligence, University of Cambridge
- **Irakli BERIDZE** – UN Centre for AI and UNICRI
- **Michelle Tuveson** – Centre for Risk Studies, University of Cambridge
- **Mark Halverson** – Founder and CEO at Human Ecology Holdings and Precision Autonomy
- **Mary Cummings, Ph.D** – Associate Professor in the Department of Mechanical Engineering and Materials Science, Duke University
- **Shahar Avin** – University of Cambridge, Centre for the Study of Existential Risk, Post-Doc
- **Niels ten Oever** – Head of Digital, Article 19, Co-chairing Research Group on Human Rights Protocol Considerations in the Internet Research Taskforce (IRTF)
- **Elizabeth D. Gibbons** – Senior Fellow and Visiting Scientist at the FXB Center for Health and Human Rights at Harvard University and a Distinguished Visiting Fellow at the Kozmetsky Center of Excellence in Global Finance at St. Edwards University
- **Alexi Grinbaum** – Researcher at CEA (French Alternative Energies and Atomic Energy Commission)

Global Initiative Membership

- **Olia Kanevskaia** – Tilburg Law and Economic Center, PhD Candidate in International Standardization; Graduate Intern at the World Trade Organization
- **Cyrus Hodes** – Director and Co-Founder, The AI Initiative; Representative for the MENA Region, The Future Society at Harvard Kennedy School
- **Gry Hasselbalch** – CoFounder, DataEthicsEU, CoAuthor, Data Ethics - The New Competitive Advantage
- **Rupak Rathore** – Principal Consultant: Strategy, Innovation Incubation and Transformation Journey Management

Global Initiative Membership

Embedding Values into Autonomous Intelligent Systems

As we design autonomous systems that we interact with on an everyday basis, technologists and engineers are left today with the daunting task of determining and imbuing certain human values and ethics onto these autonomous and intelligent systems. In addition, a set of values identified for the design of systems for one user context or user group often does not translate directly to the design of another system, making the process of developing human aligned autonomous systems a challenge. This Committee focuses on addressing the challenges of identifying, prioritizing and imbuing human values into autonomous and intelligent systems such that we can better advance technology for humanity.

- **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
- **Raja Chatila** – 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
- **Malo Bourgon** – COO, Machine Intelligence Research Institute
- **Vanessa Evers** – Professor, Human-Machine Interaction, and Science Director, DesignLab, University of Twente
- **Alan Mackworth** – Professor of Computer Science, University of British Columbia; Former President, AAAI; Co-author of "Artificial Intelligence: Foundations of Computational Agents"
- **Laurel Riek** – Associate Professor, Computer Science and Engineering, University of California San Diego
- **Alan Winfield** – Professor, Bristol Robotics Laboratory, University of the West of England; Visiting Professor, University of York
- **Wendell Wallach** – Consultant, ethicist, and scholar, Yale University's Interdisciplinary Center for Bioethics
- **Mike Van der Loos** – Associate Prof., Dept. of Mechanical Engineering, Director of Robotics for Rehabilitation, Exercise and Assessment in Collaborative Healthcare (RREACH) Lab, and Associate Director of CARIS Lab, University of British Columbia
- **Brenda Leong** – Senior Counsel, Director of Operations, The Future of Privacy Forum
- **Francesca Rossi** – (Co-Chair) Full Professor, computer science at the University of Padova, Italy, currently at the IBM Research Center at Yorktown Heights, NY

Global Initiative Membership

- **Karolina Zawieska** – SMARTlab, University College Dublin (UCD), Ireland and Industrial Research Institute for Automation and Measurements (PIAP), Poland
- **Virginia Dignum** – Associate Professor, Faculty of Technology Policy and Management, TU Delft
- **Edson Prestes** – Professor, Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Brazil; Head, Phi Robotics Research Group, UFRGS; CNPq Fellow
- **John P. Sullins** – Professor of Philosophy, Chair of the Center for Ethics Law and Society (CELS), Sonoma State University
- **Laurence Devillers** – Professor of Computer Sciences, University Paris Sorbonne, LIMSI-CNRS 'Affective and social dimensions in spoken interactions' - member of the French Commission on the Ethics of Research in Digital Sciences and Technologies (CERNA)
- **Leanne Seeto** – Strategy and Operations at Human Ecology Holdings and Precision Autonomy
- **Sara Jordan** – Assistant Professor of Public Administration in the Center for Public Administration & Policy at Virginia Tech
- **Pablo Noriega** – Scientist, Artificial Intelligence Research Institute of the Spanish National Research Council (IIIA-CSIC), Barcelona
- **Catholijn Jonker** – Full professor of Interactive Intelligence at the Faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology
- **Nell Watson** – FRSA FIAP– Faculty AI & Robotics, Singularity University; Co-Founder of OpenEth.org; Senior Scientific Advisor to The AI Initiative at The Future Society at Harvard Kennedy School; Advisor to The Future of Sentience Society, University of Cambridge. {How to Imbue}
- **Bertram Malle** – Professor, Department of Cognitive, Linguistic, and Psychological Sciences, Co-Director of the Humanity-Centered Robotics Initiative, Brown University
- **Stephen Cave** – Executive Director of the Leverhulme Centre for the Future of Intelligence, University of Cambridge
- **Ebru Dogan** – Research Engineer VEDECOM
- **Jaan Tallinn** – Founding engineer of Skype and Kazaa; co-founder of the Future of Life Institute

Global Initiative Membership

Methodologies to Guide Ethical Research and Design

This Committee is focusing on identifying the specific tools and practices that will bring applied ethics methodologies to the workplace and any design process.

- **Raja Chatila** (Co-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
- **John C. Havens** – Executive Director, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems; Author, Heartificial Intelligence: Embracing Our Humanity to Maximize Machines (Founding Committee Co-Chair)
- **Jared Bielb**y – Co-chair, International Center for Information Ethics
- **Tim Hwang** – Fellow, Data & Society
- **Jason Milla**r – Professor, robot ethics at Carleton University
- **Sarah Spiekermann** (Co-Chair) – Chair of the Institute for Management Information Systems at Vienna University of Economics and Business; Author of Ethical IT-Innovation and Blogger on The Ethical Machine
- **Tom Guarriello, Ph.D.** – Founding Faculty member in the Master's in Branding program at New York City's School of Visual Arts, Host of RoboPsyc
- **Corinne J.N. Cath** – (Co-Chair) PhD student at The University of Oxford, Programme Officer at ARTICLE 19
- **Thomas Arnold** – Research Associate at Tufts University Human-Robot Interaction Laboratory
- **Pamela Pavliscak** – Founder, Change Sciences
- **Illah Nourbakhsh** – Professor of Robotics, The Robotics Institute, Carnegie Mellon University
- **Shannon Vallor** – William J. Rewak Professor in the Department of Philosophy at Santa Clara University in Silicon Valley; President of the international Society for Philosophy and Technology (SPT) and Executive Board member of the Foundation for Responsible Robotics
- **Sara Jordan** – Assistant Professor of Public Administration in the Center for Public Administration & Policy at Virginia Tech
- **Björn Niehaves** – Chair of Business computer science at University of Siegen

Global Initiative Membership

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

This Committee is focusing on issues related to Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI). AGI generally refers to a computer or system that can perform tasks at the same level of a human or better in multiple arenas, which raises concerns about control leakage, value alignment, and system self-improvement. This Committee is discussing issues surrounding how to build systems, today and in the near-mid future that can be designed in ways that foster safety and beneficence for society while also advancing innovation for AI and autonomous technology.

- **Malo Bourgon** (Co-Chair) – COO, Machine Intelligence Research Institute
- **Richard Mallah** (Co-Chair) – Director of Advanced Analytics, Cambridge Semantics; Director of AI Projects, Future of Life Institute
- **Paul Christiano** – PhD Student, Theory of Computing Group, UC Berkeley
- **Bart Selman** – Professor of Computer Science, Cornell University
- **Carrick Flynn** – Research Assistant at Future of Humanity Institute, University of Oxford
- **Roman Yampolskiy, PhD** – Associate Professor and Director, Cyber Security Laboratory; Computer Engineering and Computer Science, University of Louisville

Global Initiative Membership

Personal Data and Individual Access Control

This Committee is focusing on issues of personal information and privacy in relation to how individual data is used in regards to autonomous and artificially intelligent systems. While it is a given that more data provided to a system will improve its ability to foster useful insights, it is also imperative that individuals have access to and can share their data in ways that protects their rights to privacy, security, and identity.

- **Michelle Finneran Dennedy** (Co-Chair) – Vice President, Chief Privacy Officer, Cisco; Author, The Privacy Engineer's Manifesto: Getting from Policy to Code to QA to Value
- **Eva Schulz-Kamm** – Head of NXP Political Affairs and Public Co-Creation (Founding Committee Co-Chair)
- **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
- **Robert-Jan Sips** – Lead, IBM Center for Advanced Studies Benelux; Europe University Program Manager, IBM
- **Dr. Zoltan Szlavik** – Researcher, IBM Center for Advanced Studies Benelux
- **Sean Bohan** – Steering Committee Member, Project VRM
- **Dr. David A. Bray** – Senior Executive & CIO for the FCC; Eisenhower Fellow to Taiwan and Australia; Harvard Visiting Executive In-Residence
- **Gry Hasselbalch** – CoFounder DataEthicsEU, CoAuthor, Data Ethics - The New Competitive Advantage
- **Emma Lindley** – Founder, Innovate Identity
- **Joseph Jerome** – Policy Counsel, Center for Democracy & Technology
- **Katryna Dow** – CEO & Founder at Meeco
- **Walter Burrough** – Co-Founder, Augmented Choice; PhD Candidate (Computer Science) - Serious Games Institute; Science teacher
- **Walter Pienciak** – Senior Manager, Strategic Programs, IEEE
- **Jean-Gabriel Ganascia** – Professor, University Pierre et Marie Curie; LIP6 Laboratory ACASA Group Leader
- **Maria Bottis** – Associate Professor, Information Law, Ionian University
- **Ariel H. Brio** – Privacy and Data Counsel at Sony PlayStation
- **Ugo Pagallo** – University of Turin Law School; Center for Transnational Legal Studies, London; NEXA Center for Internet & Society, Politecnico of Turin

Global Initiative Membership

- **Danny W. Devriendt** – Managing director of Mediabrands Publishing (IPG) in Brussels, and the CEO of the Eye of Horus, a global think-tank for communication-technology related topics
- **Sofia C. Olhede** – Professor of Statistics and an Honorary Professor of Computer Science at University College London, London, U.K; Member of the Programme Committee of the International Centre for Mathematical Sciences
- **Dr. Louise Dennis** – Post-Doctoral Researcher in the Autonomy and Verification Laboratory at the University of Liverpool
- **Ajay Bawa** – Technology Innovation Lead, Avanade Inc.

Global Initiative Membership

Reframing Autonomous Weapons Systems

A central area of global concern around autonomous technology is potential application to physical weapons. This Committee is focused on discussing methodologies and tools to ensure that issues of accountability, human-control/intervention, and overall societal safety are adequately addressed in contemporary deliberations.

- **Richard Mallah** (Chair) – Director of Advanced Analytics, Cambridge Semantics; Director of AI Projects, Future of Life Institute
- **Peter Asaro** – Assistant Professor of Media Studies, The New School
- **Ryan Gariepy** – CTO/Co-Founder, Clearpath; Director of the Board, Open Source Robotics Foundation
- **Heather Roff, Ph.D** – Senior Research Fellow, Department of Politics and International Relations, University of Oxford; Research Scientist, Global Security Initiative, Arizona State University; Fellow, New America Foundation, Cybersecurity Initiative
- **Stuart Russell** – Professor of Computer Science, University of California, Berkeley Bernhard Schölkopf – Director, Department of Empirical Inference, Max Planck Institute for Intelligent Systems
- **Noel Sharkey** – Professor of AI and Robotics, University of Sheffield; Leverhulme Research Fellow on battlefield robots
- **Eric Horvitz** – Technical Fellow, Microsoft Research
- **Catherine Tessier** – Researcher at ONERA, France, and professor at ISAE-SUPAERO

Global Initiative Membership

Economics/Humanitarian Issues

Evolution of technology doesn't happen in isolation – rapid advances in machine automation are changing the nature of how humans work and how many jobs may be available in the future. This Committee is focusing on how to leverage these powerful emerging technologies while ensuring their benefits can be evenly distributed throughout society with an overall positive impact on humanity.

- **Raj Madhavan (Chair)** – Founder & CEO of Humanitarian Robotics Technologies, LLC, Maryland, U.S.A.
- **William Hoffman** – Associate director and head of Data-Driven Development, The World Economic Forum
- **Renaud Champion** – Founder of Robolution Capital & CEO of Primnext; Executive Director of euRobotics aisbl
- **Scott L. David** – Director of Policy at University of Washington - Center for Information Assurance and Cybersecurity
- **Yves-Alexandre de Montjoye** – Lecturer (eq. Assistant Professor), Imperial College London, Dept. of Computing and Data Science Institute
- **Rose Shuman** – Partner at BrightFront Group & Founder, Question Box
- **Hruy Tsegaye** – One of the founders of iCog Labs; a pioneer company in East Africa to work on Research and Development of Artificial General Intelligence, Ethiopia
- **Ronald C. Arkin** – Regents' Professor & Director of the Mobile Robot Laboratory; Associate Dean for Research & Space Planning, College of Computing, Georgia Institute of Technology
- **Joanna Bryson** – Visiting Research Collaborator and Visiting Fellow, Center for Information Technology Policy, Princeton University; Associate Professor, University of Bath, Intelligent Systems Research Group, Department of Computer Science

Global Initiative Membership

Law

This Committee is focusing on the legal issues related to the design and use of autonomous and intelligent systems.

- **Kay Firth-Butterfield** (Co-Chair) – Executive Director, AI Austin; Consortium on Law and Ethics of Artificial Intelligence and Robotics, Strauss Center, University of Texas; University of Texas Law School
- **Derek Jinks** (Co-Chair) – University of Texas Law School; Consortium on Law and Ethics of Artificial Intelligence and Robotics, Strauss Center, University of Texas
- **Tom D. Grant** – Fellow, Wolfson College; Senior Associate of the Lauterpacht Centre for International Law, University of Cambridge, UK
- **Andrew Woods** – Assistant Professor of Law, University of Kentucky
- **Gary Marchant** – Regents' Professor of Law, Lincoln Professor of Emerging Technologies, Law and Ethics, Arizona State University
- **Tom Burri** – Assistant Professor of International and European Law, University of St. Gallen (HSG)
- **John Frank Weaver** – Lawyer, McLane Middleton, P.A, Contributing Writer for Slate, Author, Robots Are People Too
- **Ryan Calo** – Assistant Professor of Law, the School of Law at the University of Washington
- **Yan Tougas** – Global Ethics & Compliance Officer, United Technologies Corporation
- **Clemens Canel** – University of Texas at Austin
- **Paul Moseley** – University of Texas School of Law
- **Danielle Keats Cirton** – Lois K. Macht Research Professor & Professor of Law, University of Maryland Carey School of Law
- **Joseph Jerome** – Policy Counsel, Center for Democracy & Technology
- **Miles Brundage** – AI Policy Research Fellow, Strategic AI Research Center, University of Oxford; PhD candidate, Human and Social Dimensions of Science and Technology, Arizona State University
- **Matthew Scherer** – Attorney and legal scholar based in Portland, Oregon, USA. Matthew runs the "Law and AI" blog and is the author of Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies
- **Deven Desai** – Associate Professor of Law and Ethics, Georgia Institute of Technology, Scheller College of Business
- **Daniel Hinkle** – State Affairs Counsel for the American Association for Justice

Global Initiative Membership

The following Committees will be providing Content/Language for The Initiative's Conference scheduled for June, 2017.

Affective Computing

This Committee addresses the impact on individuals and society that autonomous and intelligent 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 and intelligent systems may lead us on to best provide solutions that increase human wellbeing in line with innovation.

- **Ronald C. Arkin** (Co-Chair) – Regents' Professor & Director of the Mobile Robot Laboratory; Associate Dean for Research & Space Planning, College of Computing, Georgia Institute of Technology
- **John C. Havens** – Executive Director, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems; Author, Heartificial Intelligence: Embracing Our Humanity to Maximize Machines
- **Rosalind Picard** – Rosalind Picard, (Sc.D, FIEEE) Professor, MIT Media Laboratory, Director of Affective Computing Research; Faculty Chair, MIT Mind+Hand+Heart; Co-founder & Chief Scientist, Empatica Inc.; Co-founder, Affectiva Inc.
- **Rafael Calvo** – Professor & ARC Future Fellow, School of Electrical and Information Engineering, The University of Sydney
- **Joanna Bryson** – (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
- **Jonathan Gratch** – Research Professor of Computer Science and Psychology, Director for Virtual Human Research, USC Institute for Creative Technologie
- **Matthias Scheutz** – Professor, Bernard M. Gordon Senior Faculty Fellow, Tufts University School of Engineering
- **Cynthia Breazeal** – Associate Professor of Media Arts and Sciences, MIT Media Lab; Founder & Chief Scientist of Jibo, Inc.
- **Edson Prestes** – Professor, Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Brazil; Head, Phi Robotics Research Group, UFRGS; CNPq Fellow

Global Initiative Membership

- **John P. Sullins** – Professor of Philosophy, Chair of the Center for Ethics Law and Society (CELS), Sonoma State University
- **Robert Sparrow** – Professor, Monash University, Australian Research Council "Future Fellow", 2010-15.
- **Laurence Devillers** – Professor of Computer Sciences, University Paris Sorbonne, LIMSI-CNRS 'Affective and social dimensions in spoken interactions' - member of the French Commission on the Ethics of Research in Digital Sciences and Technologies (CERNA)
- **Joost Broekens** – Assistant Professor Affective Computing, Interactive Intelligence group; Department of Intelligent Systems, Delft University of Technology
- **Genevieve Bell** – Intel Senior Fellow Vice President, Corporate Strategy Office Corporate Sensing and Insights
- **Mark Halverson** – Founder and CEO at Human Ecology Holdings and Precision Autonomy
- **Noreen Herzfeld** – Reuter Professor of Science and Religion, St. John's University
- **Bjoern Niehaves** – Professor, Chair of Information Systems, University of Siegen

Global Initiative Membership

Classical Ethics in Information & Communication Technologies

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

- **Jared Bielby** – Co-chair, International Center for Information Ethics
- **Rafael Capurro** – Founder, International Center for Information Ethics
- **Bendert Zevenbergen** – Oxford Internet Institute, University of Oxford, Creator of the Networked Systems Ethics Guidelines
- **Katie Shilton** – Leader, Ethic & Values in Design Lab at the University of Maryland, College of Information Studies, Director of the Center for the Advanced Study of Communities and Information
- **Kai Kimppa** – Postdoctoral Researcher, Information Systems, Turku School of Economics, University of Turku
- **Rachel Fischer** – Research Officer: African Centre of Excellence for Information Ethics, Information Science Department, University of Pretoria, South Africa.
- **Soraj Hongladarom** – President at The Philosophy and Religion Society of Thailand
- **Pak-Hang Wong** – Lecturer, Department of Social Science, Hang Seng Management College, Hong Kong.
- **Oliver Bendel** – Professor of Information Systems, Information Ethics and Machine Ethics, University of Applied Sciences and Arts Northwestern Switzerland FHNW
- **Miguel Á. Pérez Álvarez** – Coord. Pedagogía (modalidad a distancia) Div. Sistema de Universidad Abierta y Educación a Distancia Facultad de Filosofía y Letras Universidad Nacional Autónoma de México
- **Dr Sara Wilford** – Senior Lecturer, Research Fellow, School of Computer Science and Informatics, Centre for Computing and Social Responsibility, De Montfort University
- **Dr Neil McBride** – Reader in IT Management, School of Computer Science and Informatics, Centre for Computing and Social Responsibility, De Montfort University
- **Dr. John Burgess** – John T. F. Burgess, PhD, STM, MLIS. Assistant Professor / DE Coordinator, School of Library and Information Studies, The University of Alabama
- **Kristene Unsworth** – Assistant Professor, The College of Computing & Informatics, Drexel University
- **Wolfgang Hofkirchner** – Associate Professor, Institute for Design and Technology Assessment, Vienna University of Technology

Global Initiative Membership

The EPIC AI/AS Committee
Effective Policymaking for Innovative Communities involving Artificial Intelligence and Autonomous Systems (EPIC AI/AS)

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.

- **Dr. David A. Bray** (Co-Chair) – Visiting Executive In-Residence at Harvard University; Eisenhower Fellow to Taiwan and Australia; Federal Communications Chief Information Officer
- **Michael Krigsman** (Co-Chair) – Internationally recognized industry analyst, writer, and host of CXOTALK
- **Anja Kaspersen** – Former Head of International Security, World Economic Forum and head of strategic engagement and new technologies at the international committee of Red Cross (ICRC)
- **Corinne J.N. Cath** – PhD student at The University of Oxford, Programme Officer at ARTICLE 19
- **Darrell M. West** – Vice President and Director, Governance Studies | Founding Director, Center for Technology Innovation | The Douglas Dillon Chair
- **John C. Havens** – Executive Director, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems; Author, Heartificial Intelligence: Embracing Our Humanity to Maximize Machines
- **Karen S. Evans** – National Director, U.S. Cyber Challenge and former Administrator for the Office of Electronic Government and Information Technology, Executive Office of the President
- **Kay Firth-Butterfield** – Executive Director, AI Austin; Consortium on Law and Ethics of Artificial Intelligence and Robotics, Strauss Center, University of Texas; University of Texas Law School
- **Dr. Konstantinos Karachalios** – Managing Director, IEEE-Standards Association
- **Manu Bhardwaj** – Senior Advisor on Technology and Internet Policy to the Under Secretary of State at U.S. Department of State

Global Initiative Membership

- **Dr. Peter S. Brooks** – Institute for Defense Analyses; Science and Technology Policy Institute
- **Stephanie Wander** – Senior Manager, Prize Development, XPRIZE
- **Evangelos Simoudis** – Co-Founder and Managing Director, Synapse Partners Author, The Big Data Opportunity in our Driverless Future
- **Carolyn Nguyen** – Director, Technology Policy at Microsoft
- **Michelle Finneran Dennedy** – Vice President, Chief Privacy Officer, Cisco; Author, The Privacy Engineer's Manifesto: Getting from Policy to Code to QA to Value
- **Philip Hall** – Member, IEEE-USA Government Relations Council

Global Initiative Membership

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/AS 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.

- **Monique Morrow** (Co-Chair) – CTO New Frontiers Engineering at Cisco
- **Jay Iorio Chair** (Co-Chair) – Director of Innovation, IEEE Standards Association
- **Leanne Seeto** – Strategy and Operations at Human Ecology Holdings and Precision Autonomy
- **Katryna Dow** – CEO & Founder at Meeco
- **Pablo Noriega** – Scientist, Artificial Intelligence Research Institute of the Spanish National Research Council (IIIA-CSIC), Barcelona
- **BC Biermann, PhD** – Founder, The Heavy Projects
- **Scott Kesselman** – Advocacy and Public Affairs, Association for Unmanned Vehicle Systems International; Co-founder, writer, director and producer of experimental theater company, Blacknote Theatre

Global Initiative Membership

The following Committees (in conjunction with the Executive Committee) provide ongoing strategic guidance for The Initiative.

The Drafting Committee

The Drafting Committee is tasked with helping take drafts of *Ethically Aligned Design* and iterating them with Committee Chairs after face-to-face meeting of The IEEE Global Initiative.

- **Kay Firth-Butterfield (Chair)** – Executive Director, AI Austin; Consortium on Law and Ethics of Artificial Intelligence and Robotics, Strauss Center, University of Texas; University of Texas Law School
- **John C. Havens** – Executive Director, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems; Author, Heartificial Intelligence: Embracing Our Humanity to Maximize Machines
- **Raja Chatila** – 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
- **Tom D. Grant** – Fellow, Wolfson College; Senior Associate of the Lauterpacht Centre for International Law, University of Cambridge, UK
- **Dr. Victoria Wang** – CEO, China IP Group
- **Deven Desai** – Associate Professor of Law and Ethics, Georgia Institute of Technology, Scheller College of Business, Atlanta, Georgia, U.S.A.
- **Francesca Rossi** – Full Professor, computer science at the University of Padova, Italy, currently at the IBM Research Center at Yorktown Heights, NY

Global Initiative Membership

The Standards Committee

The Standards Committee is designed to inform The Executive Committee and Initiative Members about ongoing global Standards and practices related to AI/AS technologies so the Program's efforts can be as timely and culturally relevant as possible. The Committee is also designed to provide educational resources to help Initiative members prepare Standards proposals for consideration for IEEE-SA where there is consensus that, 1) Their Committee feels there is an issue that could benefit by standardization, 2) It has been determined there are currently no Standards related to this issue, and 3) The Committee's Standard Proposal has been written in a way that will provide it the highest level of being accepted where possible.

- **Alan Winfield** (Co-Chair) – Professor, Bristol Robotics Laboratory, University of the West of England; Visiting Professor, University of York
- **John C. Havens** – Executive Director, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems; Author, Heartificial Intelligence: Embracing Our Humanity to Maximize Machines
- **Don Wright** – President, Standards Strategies, LLC; 2016 IEEE Standards Association President-Elect
- **David Alan Grier** – Principal, Technologies Practice at Djaghe, LLC; Associate Professor of International Science and Technology Policy, Center for International Science & Technology Policy; Host/Creator of, "How We Manage Stuff" podcast
- **Laurence Devillers** – Professor of Computer Sciences, University Paris Sorbonne, LIMSI-CNRS 'Affective and social dimensions in spoken interactions' - member of the French Commission on the Ethics of Research in Digital Sciences and Technologies (CERNA)
- **Sara Jordan** – Assistant Professor of Public Administration in the Center for Public Administration & Policy at Virginia Tech

Global Initiative Membership

Ecosystem Mapping Committee

The Ecosystem mapping Committee is designed to help support the overarching, ongoing efforts of The Initiative by providing research regarding the general landscape of AI/AS global technologies.

- **Stephanie Wander** (Chair) – Senior Manager, Prize Development, XPRIZE
- **Cherry Tom** – Emerging Technologies Intelligence Manager at IEEE
- **Paula Boddington** – Senior Researcher, Department of Computer Science, University of Oxford
- **Jim Isaak** – President Emeritus, IEEE Computer Society; 2003/4 IEEE Director; 2015 VP IEEE Society on Social Implications of Technology
- **Mark A. Vasquez** – Strategic Program Development Sr. Manager, Meetings, Conferences & Events for IEEE

Global Initiative Membership

The "Lexonomy" Committee

The "Lexonomy" ("An Illustrated Lexicon of Autonomy") Committee is focused on identifying common terms, issues, and themes across Committees concerns and candidate recommendations. The goal of this work is to determine where groups share common intent, values, and context, but may be utilizing different terminology.

Consolidating terms within a context allows for more precise and broad communication of the recommendations of each committee. The group intends to test the consolidated terms against cultural, language, generational, and other considerations to ensure the messaging from committees can create impact globally. The group is building a baseline of terminology that can be communicated through multiple engaging, immersive, and emotive vehicles to support the human-aligned adoption of Autonomous and Intelligent capabilities.

- **Mark Halverson** (Chair) – Founder and CEO at Human Ecology Holdings and Precision Autonomy
- **Leanne Seeto** – Strategy and Operations at Human Ecology Holdings and Precision Autonomy
- **Scott Kesselman** – Advocacy and Public Affairs, Association for Unmanned Vehicle Systems International; Co-founder, writer, director and producer of experimental theater company, Blacknote Theatre
- **Dr. Craig A. Lindley** – Senior Principal Research Scientist, Decision Sciences Program, CSIRO Data61
- **Richard Bartley** – Security Senior Principal, Emerging Technology Security Group, Accenture

Global Initiative Membership

Special Thanks to Attendees of our SEAS Europe Event that Contributed to the Iteration of our document, *Ethically Aligned Design*:

Ronald Arkin, Thomas Arnold, Peter Asaro, Shahar Avin, Elizabeth Barnes, Dr. Shima Beigi, Irakli Beridze, Jordi Bieger, Paula Boddington, Malo Bourgon, Miles Brundage, Joanna Bryson, Thomas Burri, Prof. Dr. Rafael Capurro, Corinne Cath, Stephen Cave, Renaud Champion, Raja Chatila, Alexander Deak, Louise Dennis, Deven Desai, Laurence Devillers, Danny Devriendt, Virginia Dignum, Ebru Dogan, Katryna Dow, Kay Firth-Butterfield, Ryan Gariepy, Tony Gillespie, Thomas Dale Grant, Mark Halverson, Sobia Hamid, Robert Hart, John C. Havens, Dirk Helbing, Cyrus Hodes, Brianna Hunt, Tim Hwang, Sara Jordan, Konstantinos Karachalios, Anja Kaspersen, Daan Kayser, Scott Kesselman, Irene Kitsara, Jun Lin, Raj Madhavan, Richard Mallah, Bertram Malle, Ajung Moon, Monique Morrow, Paul Moseley, Clara Neppel, Carolyn Nguyen, Bjoern Niehaves, Pablo Noriega, Sean O Heigeartaigh, Sofia Olhede, Brian O'Rourke, Julian Padget, Ugo Pagallo, Jerome Perrin, Walter Pienciak, Edson Prestes, Huw Price, Heather Roff, Francesca Rossi, Stuart Russell, Anders Sandberg, Filippo Santoni De Sio, Daniel Saxon, Matthias Scheutz, Bernhard Schölkopf, Leanne Seeto, Andrew Snyder-Beattie, Sarah Spiekermann, Bas Steunebrink, Suhas Subramanyam, John P. Sullins, Zoltán Szlávik, Jaan Tallinn, Valerio Targon, Niels Ten Oever, Catherine Tessier, Cherry Tom, Aimee Van Wynsberghe, Harko Verhagen, Wendell Wallach, Toby Walsh, Axel Walz, Victoria Wang, Alan Winfield, Bendert Zevenbergen.