Version 2 - For Public Discussion

IEEE
Advancing Technology for Humanity

ETHICALLY ALIGNED DESIGN

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

Ethically Aligned Design – Version 2 Request for Input

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

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

The document's purpose is to:

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

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

Input about *Ethically Aligned Design* should be sent by email no later than 12 March 2018 and will be made publicly available at the website of *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems* no later than 30 April 2018. Details on how to submit public comments are available via our Submission Guidelines.

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

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

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Table of Contents

Executive Summary

Introduction	2
The Mission of The IEEE Global Initiative	3-4
Who We Are	5
Ethically Aligned Design, v2 (Overview)	6-9
Our Process	10-12
How to Cite *Ethically Aligned Design*	13
Our Appreciation	14-16
Disclaimers	17-19

Committees featured in EADv1 (with updated content)

General Principles	20-32
Embedding Values into Autonomous Intelligent Systems	33-54
Methodologies to Guide Ethical Research and Design	55-72
Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI)	73-82
Personal Data and Individual Access Control	83-112
Reframing Autonomous Weapons Systems	113-130
Economics/Humanitarian Issues	131-145
Law	146-161

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Table of Contents

New Committees for EADv2 (with new content)

Affective Computing	162-181
Policy	182-192
Classical Ethics in A/IS	193-216
Mixed Reality in ICT	217-239
Well-being	240-263

Important Links

- Website of The IEEE Global Initiative
- Full Listing of The IEEE Global Initiative Membership
- Ethically Aligned Design Version 1

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

Introduction

As the use and impact of autonomous and intelligent systems (A/IS) become pervasive, we need to establish societal and policy guidelines in order for such systems to remain human-centric, serving humanity's values and ethical principles. These systems 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 people and technology that is needed for its fruitful, pervasive use in our daily lives.

To be able to contribute in a positive, non-dogmatic way, we, the techno-scientific communities, need to enhance our self-reflection, we need to have an open and honest debate around our imaginary, our sets of explicit or implicit values, our institutions, symbols and representations.

Eudaimonia, as elucidated by Aristotle, is a practice that defines human well-being 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.

Whether our ethical practices are Western (Aristotelian, Kantian), Eastern (Shinto, Confucian), African (Ubuntu), or from a different tradition, by creating autonomous and intelligent systems that explicitly honor inalienable human rights and the beneficial values of their users, we can prioritize the increase of human well-being as our metric for progress in the algorithmic age. Measuring and honoring the potential of holistic economic prosperity should become more important than pursuing one-dimensional goals like productivity increase or GDP growth.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

The Mission of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

To ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.

By "*stakeholder*" we mean anyone involved in the research, design, manufacture, or messaging around intelligent and autonomous systems, including universities, organizations, governments, and corporations making these technologies a reality for society.

Our goal is that *Ethically Aligned Design* will provide insights and recommendations that provide a key reference for the work of technologists in the related fields of science and technology in the coming years. To achieve this goal, in the current version of *Ethically Aligned Design* (EAD2v2), we identify pertinent "Issues" and "Candidate Recommendations" we hope will facilitate the emergence of national and global policies that align with these principles.

The IEEE Global Initiative brings together several hundred participants from six continents, who are thought leaders from academia, industry, civil society, policy and government in the related technical and humanistic disciplines to identify and find consensus on timely issues.

A second goal of The IEEE Global Initiative is to provide recommendations for IEEE Standards based on *Ethically Aligned Design*. *Ethically Aligned Design* (v1 and v2) and members of The IEEE Global Initiative are the inspiration behind the suite of IEEE P7000™ Standards Working Groups that are free and open for anyone to join.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

For more information or to join any Working Group, please click on the links below:

IEEE P7000™ - Model Process for Addressing Ethical Concerns During System Design

IEEE P7001™ - Transparency of Autonomous Systems

IEEE P7002™ - Data Privacy Process

IEEE P7003™ - Algorithmic Bias Considerations

IEEE P7004™ - Standard on Child and Student Data Governance

IEEE P7005™ - Standard for Transparent Employer Data Governance

IEEE P7006™ - Standard for Personal Data Artificial Intelligence (AI) Agent

IEEE P7007™ - Ontological Standard for Ethically Driven Robotics and Automation Systems

IEEE P7008™ - Standard for Ethically Driven Nudging for Robotic, Intelligent, and Automation Systems

IEEE P7009™ - Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems

IEEE P7010™ - Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

Who We Are

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems ("The IEEE Global Initiative") is a program of The Institute of Electrical and Electronics Engineers ("IEEE"), the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity with over 420,000 members in more than 160 countries.

The IEEE Global Initiative provides the opportunity to bring together multiple voices in the related technological and scientific communities to identify and find consensus on timely issues.

IEEE will make all versions of *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 at IEEE to foster open, broad, and inclusive conversation about ethics in technology, known as the IEEE TechEthics™ program.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

Ethically Aligned Design v2 — Overview

I. Purpose

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

II. Goals

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

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

III. Objectives

Personal Data Rights and Individual Access Control

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

Well-being Promoted by Economic Effects

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

Legal Frameworks for Accountability

The convergence of intelligent systems and robotics technologies has led to the development of systems with attributes that simulate those of human beings in terms of partial autonomy, ability to perform specific intellectual tasks, and may even have a human physical appearance. The issue of the legal status of complex intelligent and autonomous technical systems thus intertwines with broader legal questions regarding how to ensure accountability and allocate liability when such systems cause harm. Some examples of general frameworks to consider include the following:

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

Transparency and Individual Rights

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

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

Policies for Education and Awareness

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

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

IV. Foundations

Classical Ethics

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

Well-being Metrics

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

Embedding Values into Autonomous Systems

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

Methodologies to Guide Ethical Research and Design

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

V. Future Technology Concerns

Reframing Autonomous Weapons

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

- Ensuring meaningful human control of weapons systems
- Designing automated weapons with audit trails to help guarantee accountability and control
- Including adaptive and learning systems that can explain their reasoning and decisions to human operators in a transparent and understandable way
- Training responsible human operators of autonomous systems who are clearly identifiable
- Achieving behavior of autonomous functions that is predictable to their operators
- Ensuring that the creators of these technologies understanding the implications of their work
- Developing professional ethical codes to appropriately address the development of autonomous systems intended to cause harm

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

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

Affective Computing

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

Mixed Reality

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

Our Process

To ensure greatest cultural relevance and intellectual rigor in our work, The IEEE Global Initiative has been globally crowdsourcing feedback for Versions 1 and 2 of *Ethically Aligned Design*.

We released *Ethically Aligned Design Version 1* as a Request for Input on December of 2016 and received over two hundred pages of in-depth feedback about the draft. As a way to highlight insights inspired by the feedback we received, Sara Mattingly-Jordan of The IEEE Global Initiative also wrote the report, Becoming a Leader in Global Ethics.

We are releasing *Ethically Aligned Design Version 2 (EADv2)* as a Request for Input once again to gain further insights about the eight original sections from EADv1, along with unique/new feedback for the five new sections included in EADv2.

Next Steps

The IEEE Global Initiative is currently creating an organizational committee composed of representatives of all our Committees and IEEE P7000™ Working Groups to do the following in order to prepare the final version of *Ethically Aligned Design* to be released in 2019:

- Create criteria for Committees to vote on all "Candidate Recommendations" becoming "Recommendations" based on the General Principles of *Ethically Aligned Design* that are in accordance with the Mission Statement of The IEEE Global Initiative. This voting process will be based on the consensus-based protocols provided by IEEE-SA.
- Create a rigorous methodology to best incorporate feedback received from EADv1 and EADv2, working to holistically consider global and diversity-based considerations for content inclusion.
- Use the glossary we have produced as a key tool for synthesizing content for final version of EAD, unifying terms as much as possible.

Final Version of *Ethically Aligned Design* – Format and Goals

The final version of *Ethically Aligned Design* will be made available in the following formats:

- **Handbook.** While specific formatting is still under consideration, the final version of *Ethically Aligned Design* will feature "Recommendations" (versus "Candidate Recommendations") for all existing and future "Issues" voted on by Members of The IEEE Global Initiative. It is very likely the final version of EAD will not be broken into sections according to Committees (as with EADv1 and EADv2) but according to themes or principles to be decided on by the organizational committee mentioned above. While not an official IEEE position statement, "Recommendations" will be

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

created to be easily utilized by technologists and policy makers focusing on autonomous and intelligent systems design, usage, and governance.

- **Educational materials.** The IEEE Global Initiative would like to convert the handbook version of *Ethically Aligned Design* into an academically oriented book/educational materials. Evergreen in nature, these would be targeted to academics, engineers, and technologists looking for global guidance to be used in university, post-grad, or other educational settings where ethics in technology or the issues EAD comprises would be taught.

Incorporating Feedback

While it was our intention to directly accept or review all feedback we received for EADv1, we were (happily) overwhelmed with the fantastic response we received. However, to most holistically include feedback from EADv1 and EADv2 into our overall process we have created a Glossary and are working to increase more global representation and diversity in our work. Specifically:

Glossary

We received a great deal of feedback on the need for aligned recommendations for key terms in *Ethically Aligned Design*. To that end, we created a Glossary Committee and launched the first draft of our Glossary at the same time we released EADv2. Our goal is to refine our Glossary so that by mid-2018, based on aggregated feedback to all sections of EAD (Versions 1 and 2), we can standardize definitions that reflect a global and holistic set of definitions to be implemented by all Committees in the final version of EAD.

More Global Representation/Diversity

We received a great deal of feedback noting that EADv1 was fairly "Western" in its cultural orientation. This makes sense, as the initial 100 members working on EADv1 where largely from North America and the European Union. Since the release of EADv1, however, we have:

- Added members from China, Korea, Japan, Brazil, Mexico, the Russian Federation, Iran, Thailand, and Israel along with new people from the United States and the European Union. In addition to the 250 members of the Initiative, there are also now more than 400 global members in the IEEE P7000™ Working Groups that EAD inspired.
- Supported the members translating the Executive Summary of EADv1 into multiple languages.
- Added our new "Classical Ethics in A/IS" Committee.
- Created the Becoming a Leader in Global Ethics report.
- Commissioned a report from our newer global members about the state of A/IS Ethics in their regions.
- Created an Outreach Committee to help identify and incorporate work being done

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

in A/IS ethics by women, people of color, students, and other groups representing the full spectrum of society that we are hoping to positively influence with our work. We are currently working with members of The Reboot Retreat, AI4ALL, and other leaders within IEEE to help us ensure that The IEEE Global Initiative and the final version of *Ethically Aligned Design* are as holistically representative and relevant as possible.

Terminology Update

There is no need to use the term *artificial intelligence* in order to conceptualize and speak of technologies and systems that are meant to extend our human intelligence or be used in robotics applications. For this reason, we use the term, *autonomous and intelligent systems* (or *A/IS*) in the course of our work. We chose to use this phrase encapsulating multiple fields (machine learning, intelligent systems engineering, robotics, etc.) throughout *Ethically Aligned Design*, Version 2 to ensure the broadest application of ethical considerations in the design of these technologies as possible.

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.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

How to Cite Ethically Aligned Design

Please cite Version 2 of *Ethically Aligned Design* in the following manner:

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, Version 2. IEEE, 2017. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

Our Appreciation

We wish to thank our Executive Committee and Chair of The IEEE Global Initiative:

Executive Committee Officers

Raja Chatila, *Chair*

Kay Firth-Butterfield, *Vice-Chair*

John C. Havens, *Executive Director*

Executive Committee Members

Dr. Greg Adamson, Ronald C. Arkin, Virginia Dignum, Danit Gal, Philip Hall, Malavika Jayaram, Sven Koenig, Raj Madhavan, Richard Mallah, Hagit Messer Yaron, AJung Moon, Monique Morrow, Francesca Rossi, Alan Winfield

Committee Chairs

- **General Principles:** Alan Winfield and Mark Halverson
- **Embedding Values into Autonomous Intelligent Systems:** Francesca Rossi and Bertram F. Malle
- **Methodologies to Guide Ethical Research and Design:** Raja Chatila and Corinne J.N. Cath
- **Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI):** Malo Bourgon and Richard Mallah
- **Personal Data and Individual Access Control:** Katryna Dow and John C. Havens
- **Reframing Autonomous Weapons Systems:** Peter Asaro
- **Economics/Humanitarian Issues:** Kay Firth-Butterfield and Raj Madhavan
- **Law:** Kay Firth-Butterfield and Derek Jinks
- **Affective Computing:** Ronald C. Arkin and Joanna J. Bryson
- **Classical Ethics in A/IS:** Jared Bielby
- **Policy:** Kay Firth-Butterfield and Philip Hall
- **Mixed Reality:** Monique Morrow and Jay Iorio
- **Well-being:** Laura Musikanski and John C. Havens
- **Drafting:** Kay Firth-Butterfield and Deven Desai
- **Industry:** Virginia Dignum and Malavika Jayaram
- **Communications:** Leanne Seeto and Mark Halverson
- **Glossary:** Sara M. Jordan
- **Outreach:** Danit Gal

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

We wish to express our appreciation for the reports, organizations, and individuals that have contributed research and insights helping to increase awareness around ethical issues in the realm of intelligent and autonomous systems, including *(but not limited to, and in no particular order)*:

Reports

The Future of Life Institute's Asilomar AI Principles, The AI Now 2017 Report, Human Rights in the Robot Age Report from The Rathenau Instituut, Report of COMEST on Robotics Ethics from UNESCO, The European Parliament's Recommendations to the Commission on Civil Law Rules on Robotics, Artificial intelligence — The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society report from the *European Economic and Social Committee* (Rapporteur: Catelijne MULLER), OECD's report, Going Digital: Making the Transformation Work for Growth and Well-Being, USACM's Statement on Algorithmic Transparency and Accountability, Guide to the Ethical Design and Application of Robots and Robotic Systems (British Standards Institute), Japan's Basic Rules for AI Research, Éthique de la Recherche en Robotique (CERNA), Charta der Digitalen Grundrechte der Europäischen Union (Charter of the Digital Fundamental Rights of the European Union), Telecommunications Research Laboratory, "AI Network Kentōkai Kaigi Hōkokusho 2016: AI Network no Eikyō to Risk — Chiren Shakai (WINS) no Jitsugen ni Muketa Kadai" (AIネットワーク化検討会議 報告書2016 の公表 -「AIネットワーク化の影響とリスク - 智連社会 (WINS(ウインズ)) の実現に向けた課題 -」) [The Conference on Networking among AIs Report (2016): Impacts and Risks of AI Networking Issues for the Realization of Wisdom Network Society, (WINS)], Japanese Ministry of Internal Affairs and Communications, The Information Technology Industry Council's AI Policy Principles, Intel's Artificial Intelligence — The Public Policy Opportunity, IEEE European Public Policy Initiative's position statement, Artificial Intelligence: Calling on Policy Makers to Take a Leading Role in Setting a Long Term AI Strategy, IEEE-USA's position statement on Artificial Intelligence Research, Development and Regulation, The IEEE Global Initiative's Prioritizing Human Well-being in the Age of Artificial Intelligence.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

Organizations

The Association for the Advancement of Artificial Intelligence and their formative work on AI Ethics, The Future of Life Institute, The Partnership on AI to Benefit People and Society, The Foundation for Responsible Robotics, AI & Society, Machine Intelligence Research Institute, The International Center for Information Ethics, The African Center of Excellence for Information Ethics, The 4TU Center for Ethics and Technology, The Center for the Study of Existential Risk, The Leverhulme Center for the Future of Intelligence, The Future of Humanity Institute, The Japanese Society for Artificial Intelligence, The Association for Computing Machinery, Future Advocacy, ACM Special Interest Group on Artificial Intelligence, The World Economic Forum's Global Future Council of Artificial Intelligence and Robotics, The Digital Asia Hub, The AI Initiative, The Open Roboethics Institute, The Dalai Lama Center for Ethics and Transformative Values at MIT, The Ethics Initiative at MIT Media Lab, The IEEE-USA Government Relations Council Artificial Intelligence Committee, The IEEE Robotics and Automation Society Committee on Robot Ethics, The IEEE Robotics and Automation Society, The IEEE Society on Social Implications of Technology, The IEEE Computer Society, The IEEE Computational Intelligence Society, The IEEE Systems, Man and Cybernetics Society, The IEEE Symbiotic Autonomous Systems Initiative.

People

We would like to warmly recognize the leadership and constant support of The IEEE Global Initiative by Dr. Ing. Konstantinos Karachalios, Managing Director of the IEEE Standards Association and a member of the IEEE Management Council.

We would especially like to thank Eileen M. Lach, the IEEE General Counsel and Chief Compliance Officer, who invested her time and expertise in fully reviewing this entire document, with the heartfelt conviction that there is a pressing need to focus the global community on highlighting ethical considerations in the development of autonomous and intelligent systems.

Special thanks to Dr. Peter S. Brooks for his contributions to the Overview of EADv2.

Thank You to Our Members and IEEE Team

Our progress and the ongoing positive influence of this work is due to the volunteer experts serving on our Committees and IEEE P7000™ Standards Working Groups, along with the IEEE staff who support our efforts. Thank you for your dedication toward defining, designing, and inspiring the ethical PRINCIPLES and STANDARDS that will ensure that intelligent and autonomous systems and the technologies associated therewith will positively benefit humanity.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

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 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 of IEEE or any other organization with which is affiliated. It is intended to be a working reference tool created in an inclusive process by those in the relevant scientific and engineering communities 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 positions taken by 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 the second 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 concurrence on content in this document because of its versioning format and the concurrence-building process of The IEEE Global Initiative. 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 concurrence-building process where members contributing content are proposing candidate recommendations so as not to imply these are final recommendations at this time.

Our Membership

The IEEE Global Initiative currently has more than 250 experts from all but one continent involved in our work, and we are eager for new voices and perspectives to join our work.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

Copyright, 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

Copyright © 2017 by The Institute of Electrical and Electronics Engineers, Incorporated

Published December 2017

Printed in the United States of America.

IEEE is a registered trademark owned by The Institute of Electrical and Electronics Engineers, Incorporated.

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

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

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

This work is made available under the Creative Commons Attribution Non-Commercial License.

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

Find IEEE standards and standards-related product listings at: 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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Executive Summary

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 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.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

General Principles

The General Principles Committee seeks to articulate high-level ethical concerns that apply to all types of autonomous and intelligent systems (A/IS*), regardless of whether they are physical robots (such as care robots or driverless cars) or software systems (such as medical diagnosis systems, intelligent personal assistants, or algorithmic chat bots). We are motivated by a desire to create ethical principles for A/IS that:

1. Embody the highest ideals of human beneficence as a superset of Human Rights.

2. Prioritize benefits to humanity and the natural environment from the use of A/IS. Note that these should not be at odds — one depends on the other. Prioritizing human well-being does not mean degrading the environment.

3. Mitigate risks and negative impacts, including misuse, as A/IS evolve as socio-technical systems. In particular by ensuring A/IS are accountable and transparent.

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

We have identified principles created by our Committee as well as aggregated principles reflected from other Committees of The IEEE Global Initiative. Therefore, readers should note that some general principles are reiterated and elaborated by other committees, as appropriate to the specific concerns of those committees. 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.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

General Principles

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.

*The acronym A/IS is shorthand for Autonomous and Intelligent Systems. When represented in this way, it refers to the overlapping concerns about the design, development, deployment, decommissioning, and adoption of autonomous or intelligent software when installed into other software and/or hardware systems that are able to exercise independent reasoning, decision-making, intention forming, and motivating skills according to self-defined principles.

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

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Principle 1 — Human Rights

Issue:

How can we ensure that A/IS do not infringe upon human rights?

Background

Human benefit is an important goal of A/IS, as is respect for human rights set out, *inter alia*, in The Universal Declaration of Human Rights, the International Covenant for Civil and Political Rights, the Convention on the Rights of the Child, Convention on the Elimination of all forms of Discrimination against Women, Convention on the Rights of Persons with Disabilities, and the Geneva Conventions. Such rights need to be fully taken into consideration by individuals, companies, professional bodies, research institutions, and governments alike to reflect the following concerns:

1. A/IS should be designed and operated in a way that both respects and fulfills human rights, freedoms, human dignity, and cultural diversity.
2. A/IS must be verifiably safe and secure throughout their operational lifetime.
3. If an A/IS causes harm it must always be possible to discover the root cause, by assuring *traceability* for said harm (see also Principle 4 — Transparency).

While their interpretation may change over time, human rights as defined by international law, provide a unilateral basis of creating any A/IS system as they affect humans, their emotions, data, or agency. While the direct coding of human rights in A/IS may be difficult or impossible based on contextual use, newer guidelines from The United Nations, such as the Ruggie principles, provide methods to pragmatically implement human rights ideals within business or corporate contexts that could be adapted for engineers and technologists. In this way technologists can take account of rights in the way A/IS are operated, tested, validated, etc. In short, human rights should be part of the ethical risk assessment of A/IS.

Candidate Recommendations

To best honor human rights, society must assure the safety and security of A/IS so that 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 assuring that the

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

General Principles

use of A/IS does not infringe upon human rights, freedoms, dignity, and privacy, and of traceability to contribute to the building of public trust in A/IS.

2. A way to translate existing and forthcoming legal obligations into informed policy and technical considerations is needed. Such a method should allow for differing cultural norms as well as legal and regulatory frameworks.

3. For the foreseeable future, A/IS should not be granted rights and privileges equal to human rights: A/IS should always be subordinate to human judgment and control.

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 A/IS to best honor human rights:

- The Universal Declaration of Human Rights, 1947.
- 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.
- 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.
- For an example of a guide on how to conduct an ethical risk assessment see British Standards Institute BS8611:2016, Guide to the Ethical Design and Application of Robots and Robotic Systems.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

General Principles

Principle 2 — Prioritizing Well-being

Issue:

Traditional metrics of prosperity do not take into account the full effect of A/IS technologies on human well-being.

Background

A focus on creating ethical and responsible AI has been increasing among technologists in the past 12 to 16 months. Key issues of transparency, accountability, and algorithmic bias are being directly addressed for the design and implementation of A/IS. While this is an encouraging trend, a key question facing technologists today is beyond designing responsible A/IS. That question is, What are the specific metrics of societal success for "ethical AI" once released to the world?

For A/IS technologies to provably advance benefit for humanity, we need to be able to define and measure the benefit we wish to increase. Avoiding negative unintended consequences and increasing value for customers and society (today measured largely by gross domestic product (GDP), profit, or consumption levels) are often the only indicators utilized in determining success for A/IS.

Well-being, for the purpose of *The IEEE Global Initiative*, is defined as encompassing human satisfaction with life and the conditions of life as well as an appropriate balance between positive and negative affect. This definition is based on the Organization for Economic Co-Operation and Development's (OECD) *Guidelines on Measuring Subjective Well-being* that notes, "Being able to measure people's quality of life is fundamental when assessing the progress of societies. There is now widespread acknowledgement that measuring subjective well-being is an essential part of measuring quality of life alongside other social and economic dimensions." Data is also currently being gathered in governments, businesses, and other institutions using scientifically valid measurements of well-being. Since modern societies are largely constituted of A/IS users, we believe these considerations to be relevant for A/IS developers.

It is widely agreed that GDP is at best incomplete, and at worst misleading, as a metric of true prosperity for society at large and A/IS technologies (as noted in *The Oxford Handbook of Well-Being and Public Policy*). Although the concerns regarding GDP reflect holistic aspects of society versus the impact of any one technology, they reflect the lack of universal usage of well-being indicators for A/IS. A/IS undoubtedly hold positive promise for society. But beyond the critical importance of designing and manufacturing these technologies in an

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

General Principles

ethically driven and responsible manner is the seminal question of determining the key performance indicators (KPIs) of their success once introduced into society.

A/IS technologies can be narrowly conceived from an ethical standpoint; be legal, profitable, and safe in their usage; and yet not positively contribute to human well-being. This means technologies created with the best intentions, but without considering well-being metrics, can still have dramatic negative consequences on people's mental health, emotions, sense of themselves, their autonomy, their ability to achieve their goals, and other dimensions of well-being.

Nonetheless, quantitative indicators of individual well-being should be introduced with caution, as they may provoke in users an automatic urge for numerical optimization. While this tendency is theoretically unavoidable, efforts should be invested in guaranteeing that it will not flatten the diversity of human experience. The A/IS using quantitative indicators for health or happiness should therefore develop and implement measures for maintaining full human autonomy of their users.

In conclusion, it is widely agreed that de facto metrics regarding safety and fiscal health do not encompass the full spectrum of well-being for individuals or society. By not elevating additional environmental and societal indicators as pillars of success for A/IS, we risk minimizing the positive and holistic impact for humanity of these technologies. Where personal, environmental, or social factors are not prioritized as highly as fiscal metrics of success, we also risk expediting negative and irreversible harms to our planet and population.

Candidate Recommendation

A/IS should prioritize human well-being as an outcome in all system designs, using the best available, and widely accepted, well-being metrics as their reference point.

Further Resources

- IEEE P7010™, *Well-being Metrics Standard for Ethical AI and Autonomous Systems*.
- The Measurement of Economic Performance and Social Progress (2009) now commonly referred to as "The Stiglitz Report," commissioned by the then President of the French Republic. From the report: "… the time is ripe for our measurement system to shift emphasis from measuring economic production to measuring people's well-being … emphasizing well-being is important because there appears to be an increasing gap between the information contained in aggregate GDP data and what counts for common people's well-being."
- Organisation for Economic Co-Operation & Development, *OECD Guidelines for Measuring Subjective Well-being*. Paris: OECD, 2013.
- Beyond GDP (European Commission) From the site: "The Beyond GDP initiative is about developing indicators that are

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

General Principles

as clear and appealing as GDP, but more inclusive of environmental and social aspects of progress."

- Global Dialogue for Happiness, part of the annual World Government Summit, February 11, 2017.
- Organization for Economic Co-Operation and Development, OECD's Better Life Index.
- New Economics Foundation, The Happy Planet Index.
- Redefining Progress, Genuine Progress Indicator.
- The International Panel on Social Progress, Social Justice, Well-Being and Economic Organization.
- Veenhoven, R. World Database of Happiness. Rotterdam, The Netherlands: Erasmus University.
- Royal Government of Bhutan. The Report of the High-Level Meeting on Wellbeing and Happiness: Defining a New Economic Paradigm. New York: The Permanent Mission of the Kingdom of Bhutan to the United Nations, 2012.
- See also Well-being Section in *Ethically Aligned Design*, Version 2.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Principle 3 — Accountability

Issue:

How can we assure that designers, manufacturers, owners, and operators of A/IS are responsible and accountable?

Background

The programming, output, and purpose of A/IS are often not discernible by the general public. Based on the cultural context, application, and use of A/IS, people and institutions need clarity around the manufacture and deployment of these systems to establish responsibility and accountability, and 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, if necessary apportion culpability among several responsible designers, manufacturers, owners, and/or operators, to avoid confusion or fear within the general public.

Note that accountability is enhanced with transparency, thus this principle is closely linked with Principle 4 — Transparency.

Candidate Recommendations

To best address issues of responsibility and accountability:

1. Legislatures/courts should clarify issues of responsibility, culpability, liability, and accountability for A/IS where possible during development and deployment (so that manufacturers and users understand their rights and obligations).

2. Designers and developers of A/IS should remain aware of, and take into account when relevant, the diversity of existing cultural norms among the groups of users of these A/IS.

3. Multi-stakeholder ecosystems should be developed to help create norms (which can mature to best practices and laws) where they do not exist because A/IS-oriented technology and their impacts are too new (including representatives of civil society, law enforcement, insurers, manufacturers, engineers, lawyers, etc.).

4. Systems for registration and record-keeping should be created so that it is always possible to find out who is legally responsible for a particular A/IS. Manufacturers/operators/

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

General Principles

owners of A/IS should register key, high-level parameters, including:

- Intended use
- Training data/training environment (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

- Shneiderman, B. "Human Responsibility for Autonomous Agents." *IEEE Intelligent Systems* 22, no. 2, (2007): 60–61.
- Matthias, A. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6, no. 3 (2004): 175–183.
- Hevelke A., and J. Nida-Rümelin. "Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis." *Science and Engineering Ethics* 21, no. 3 (2015): 619–630.
- An example of good practice (in relation to Candidate Recommendation #3) can be found in Sciencewise – the U.K. national center for public dialogue in policy-making involving science and technology issues.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Principle 4 — Transparency

Issue:

How can we ensure that A/IS 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, transparent A/IS are ones in which it is possible to discover how and why a system made a particular decision, or in the case of a robot, acted the way it did. Note that here the term transparency also addresses the concepts of traceability, explicability, and interpretability.

A/IS will be performing tasks that are far more complex and have more effect on our world than prior generations of technology. This reality will be particularly acute with systems that interact with the physical world, thus raising the potential level of harm that such a system could cause. For example, some A/IS already have real consequences to human safety or well-being, 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 A/IS technology 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. 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 (see Principle 3— Accountability).

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

1. For users, transparency is important because it provides a simple way for them to understand what the system is doing and why.
2. For validation and certification of an A/IS, transparency is important because it exposes the system's processes and input data to 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.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

General Principles

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 to build public confidence in the technology, promote safer practices, and facilitate wider societal adoption.

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 A/IS and how they have been verified, and 3) for accident investigators, secure storage of sensor and internal state data, comparable to a flight data recorder or black box.)

*Note that IEEE Standards Working Group P7001™ has been set up in response to this recommendation.

Further Resources

- Cappelli, C., P. Engiel, R. Mendes de Araujo, and J. C. Sampaio do Prado Leite. "Managing Transparency Guided by a Maturity Model." *3rd Global Conference on Transparency Research* 1 no. 3, 1–17. Jouy-en-Josas, France: HEC Paris, 2013.

- Sampaio do Prado Leite, J. C., and C. Cappelli. "Software Transparency." *Business & Information Systems Engineering* 2, no. 3 (2010): 127–139.

- Winfield, A., and M. Jirotka. "The Case for an Ethical Black Box." *Lecture Notes in Artificial Intelligence* 10454, (2017): 262–273.

- Wortham, R. R., A. Theodorou, and J. J. Bryson. "What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems." *IJCAI-2016 Ethics for Artificial Intelligence Workshop*. New York, 2016.

- Machine Intelligence Research Institute. "Transparency in Safety-Critical Systems." August 25, 2013.

- Scherer, M. "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." *Harvard Journal of Law & Technology* 29, no. 2 (2015).

- U.K. House of Commons. "Decision Making Transparency" pp. 17–18 in Report of the U.K. House of Commons Science and Technology Committee on Robotics and Artificial Intelligence, September 13, 2016.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Principle 5 — A/IS Technology Misuse and Awareness of It

Issue:

How can we extend the benefits and minimize the risks of A/IS technology being misused?

Background

New technologies give rise to greater risk of misuse, and this is especially true for A/IS. A/IS increases the impact of risks such as hacking, the misuse of personal data, "gaming," or exploitation (e.g., of vulnerable users by unscrupulous parties). These are not theoretical risks. Cases of A/IS hacking have already been widely reported, of driverless cars for example. The EU's General Data Protection Regulation (GDPR) provides measures to remedy the misuse of personal data. The Microsoft Tay AI chatbot was famously gamed when it mimicked deliberately offensive users. 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 A/IS.

Responsible innovation requires designers to anticipate, reflect, and engage with users of A/IS thus, through education and awareness, citizens, lawyers, governments, etc. have a role to play in developing accountability structures (Principle 3). They also have a role to play in guiding new technology proactively toward beneficial ends.

Candidate Recommendations

Raise public awareness around the issues of potential A/IS technology 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 A/IS (e.g., by providing "data privacy" warnings that some smart devices will collect their user's personal data).

2. Delivering this education in scalable and effective ways, beginning with those having the greatest credibility and impact that also minimize generalized (e.g., non-productive) fear about A/IS (e.g., via credible research institutions or think tanks via social media such as Facebook or YouTube).

3. Educating government, lawmakers, and enforcement agencies 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 A/IS).

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

General Principles

Further Resources

- Greenberg, A. "Hackers Fool Tesla S's Autopilot to Hide and Spoof Obstacles." Wired, August 2016.
- (In relation to Candidate Recommendation #2) Wilkinson, C., and E. Weitkamp. *Creative Research Communication: Theory and Practice*. Manchester, UK: Manchester University Press, 2016.
- Engineering and Physical Sciences Research Council. Anticipate, Reflect, Engage and Act (AREA) Framework for Responsible Research and Innovation.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

Society has not established universal standards or guidelines for embedding human norms and values into autonomous and intelligent systems (A/IS) today. But as these systems are instilled with increasing autonomy in making decisions and manipulating their environment, it is essential they be designed to adopt, learn, and follow the norms and values of the community they serve. Moreover, their actions must be transparent in signaling their norm compliance and, if needed, they must be able to explain their actions. This is essential if humans are to develop levels of trust in A/IS that are appropriate in the specific contexts and roles in which A/IS function.

The conceptual complexities surrounding what "values" are (e.g., Hitlin and Piliavin, 2004; Malle and Dickert, 2007; Rohan, 2000; Sommer, 2016) make it currently difficult to envision A/IS that have computational structures directly corresponding to social or cultural values (such as "security," "autonomy," or "fairness"). However, it is a more realistic goal to embed explicit norms into such systems because norms can be considered instructions to act in defined ways in defined contexts, for a specific community (from family to town to country and beyond). A community's network of norms is likely to reflect the community's values, and A/IS 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 per se. (For discussion of specific values that are critical for ethical considerations of A/IS, see the sections "Personal Data and Individual Access Control" and "Well-being".)

Norms are typically expressed in terms of obligations and prohibitions, and these can be expressed computationally (e.g., Malle, Scheutz, and Austerweil, 2017; Vázquez-Salceda, Aldewereld, Dignum, 2004). At this level, norms are typically qualitative in nature (e.g., do not stand too close to people). However, the implementation of norms also has a quantitative component (the measurement of the physical distance we mean by "too close"), and the possible instantiations of the quantitative component technically enable the qualitative norm.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

To address the broad objective of embedding norms and, by implication, values into these systems, our Committee has defined three more concrete goals as described in the following sections:

1. Identifying the norms of a specific community in which A/IS operate.

2. Computationally implementing the norms of that community within the A/IS.

3. Evaluating whether the implementation of the identified norms in the A/IS are indeed conforming to the norms reflective of that community.

Pursuing these three goals represents an iterative process that is sensitive to the purpose of A/IS and their users within a specific community. It is understood that there may be clashes of values and norms when identifying, implementing, and evaluating these systems. Such clashes are a natural part of the dynamically changing and renegotiated norm systems of any community. As a result, we advocate for an approach where systems are designed to provide transparent signals (such as explanations or inspection capabilities) about the specific nature of their behavior to the individuals in the community they serve.

References

- Hitlin, S., and J. A. Piliavin. "Values: Reviving a Dormant Concept." *Annual Review of Sociology* 30 (2004): 359–393.

- Malle, B. F., and S. Dickert. "Values," *The Encyclopedia of Social Psychology*, edited by R. F. Baumeister and K. D. Vohs, Thousand Oaks, CA: Sage, 2007.

- Malle, B. F., M. Scheutz, and J. L. Austerweil. "Networks of Social and Moral Norms in Human and Robot Agents," in *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, edited by M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar, and G. S. Virk, 3–17. Cham, Switzerland: Springer International Publishing, 2017.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

- Rohan, M. J. "A Rose by Any Name? The Values Construct." *Personality and Social Psychology Review 4* (2000): 255–277.
- Sommer, A. U. *Werte: Warum Man Sie Braucht, Obwohl es Sie Nicht Gibt.* [Values. Why We Need Them Even Though They Don't Exist.] Stuttgart, Germany: J. B. Metzler, 2016.
- Vázquez-Salceda J., H. Aldewereld, and F. Dignum. "Implementing Norms in *Multiagent Systems," in Multiagent System Technologies. MATES 2004*, edited by G. Lindemann, J. Denzinger, I. J. Timm, and R. Unland. (Lecture Notes in Computer Science, vol. 3187.) Berlin: Springer, 2004.

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 1 — Identifying Norms for Autonomous Intelligent Systems

We identify three issues that must be addressed in the attempt to identify norms (and thereby values) for A/IS. The first issue asks which norms should be identified, and with which properties. Here we highlight context specificity as a fundamental property of norms. Second, we emphasize another fundamental property of norms: their dynamically changing nature, which requires A/IS to have the capacity to update their norms and learn new ones. Third, we address the challenge of norm conflicts that naturally arise in a complex social world. Resolving such conflicts requires priority structures among norms, which help determine whether, in a given context, adhering to one norm is more important than adhering to another norm.

Issue 1:
Which norms should be identified?

Background and Analysis

If machines engage in 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. But which norms? Laws are publicly documented and therefore easy to identify, so they will certainly have to be incorporated into A/IS. Social and moral norms are more difficult to ascertain, as they are expressed through behavior, language, customs, cultural symbols, and artifacts. Most important, communities (from families to whole nations) differ to various degrees in the laws and norms they follow. Therefore, generating a universal set of norms that applies to all autonomous systems is not realistic, but neither is it advisable to completely personalize an A/IS to individual preferences. However, we believe that identifying broadly observed norms of a particular community is feasible.

The difficulty of generating a set of universal norms is not inconsistent with the goal of seeking agreement over Universal Human Rights (see "General Principles" section). However, such universal rights would not be sufficient for devising an A/IS that obeys the specific norms of its community. Universal rights must, however, constrain the kinds of norms that are implemented in an A/IS.

Embedding norms in A/IS requires a clear delineation of the community in which the A/IS are to be deployed. Further, even within a particular community, different types of A/IS will demand different sets of norms.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

The relevant norms for self-driving vehicles, for example, will differ greatly from those for robots used in healthcare. Thus, we recommend that to develop A/IS capable of following social and moral norms, the first step is to identify the norms of the specific community in which the A/IS are to be deployed and, in particular, norms relevant to the kinds of tasks that the A/IS are designed to perform. Even when designating a narrowly defined community (e.g., a nursing home; an apartment complex; a company), there will be variations in the norms that apply. The identification process must heed such variation and ensure that the identified norms are representative not only of the dominant subgroup in the community but also of vulnerable and underrepresented groups.

The most narrowly defined community is a single person, and A/IS may well have to adapt to the unique norms of a given individual, such as norms of arranging a disabled person's home to accommodate certain physical limitations. However, unique individual norms must not violate norms in the larger community. Whereas the arrangement of someone's kitchen or the frequency with which a care robot checks in with a patient can be personalized without violating any community norms, encouraging the robot to use derogatory language to talk about certain social groups does violate such norms. (In the next section we discuss how A/IS might handle such norm conflicts.)

We should note that the norms that apply to humans may not always be identical to the norms that would apply to an A/IS in the same context. Empirical research involving multiple disciplines and multiple methods (see the Further Resources section) should therefore (a) investigate and document both community- and task-specific norms that apply to humans and (b) consider possible differences for A/IS deployed in these contexts. The set of empirically identified norms applicable to A/IS should then be made available for designers to implement.

Candidate Recommendation

To develop A/IS capable of following social and moral norms, the first step is to identify the norms of the specific community in which the A/IS are to be deployed and, in particular, norms relevant to the kinds of tasks that the A/IS are designed to perform.

Further Resources

- Bendel, O. *Die Moral in der Maschine: Beiträge zu Roboter- und Maschinenethik*. Hannover, Germany: Heise Medien, 2016. Accessible popular-science contributions to philosophical issues and technical implementations of machine ethics.
- Burks, S. V., and E. L. Krupka. "A Multimethod Approach to Identifying Norms and Normative Expectations within a Corporate Hierarchy: Evidence from the Financial Services Industry." *Management Science* 58 (2012): 203–217. Illustrates surveys and incentivized coordination games as methods to elicit norms in a large financial services firm.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

- Friedman, B., P. H. Kahn, A. Borning, and A. Huldtgren. "Value Sensitive Design and Information Systems," in *Early Engagement and New Technologies: Opening up the Laboratory* (Vol. 16), edited by N. Doorn, D. Schuurbiers, I. van de Poel, and M. E. Gorman, 55–95. Dordrecht: Springer, 2013. A comprehensive introduction into Value Sensitive Design and three sample applications.
- Mackie, G., F. Moneti, E. Denny, and H. Shakya. What Are Social Norms? How Are They Measured? UNICEF Working Paper. University of California at San Diego: UNICEF, 2012. A broad survey of conceptual and measurement questions regarding social norms.
- Malle, B. F. "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology* 18, no. 4 (2016): 243–256. Discusses how a robot's norm capacity fits in the larger vision of a robot with moral competence.
- Miller, K. W., M. J. Wolf, and F. Grodzinsky. "This 'Ethical Trap' Is for Roboticists, Not Robots: On the Issue of Artificial Agent Ethical Decision-Making." *Science and Engineering Ethics* 23 (2017): 389–401. This article raises doubts about the possibility of imbuing artificial agents with morality, or claiming to have done so.
- Rizzo, A., and L. L. Swisher. "Comparing the Stewart–Sprinthall Management Survey and the Defining Issues Test-2 as Measures of Moral Reasoning in Public Administration." *Journal of Public Administration Research and Theory* 14 (2004): 335–348. Describes two assessment instruments of moral reasoning (including norm maintenance) based on Kohlberg's theory of moral development.
- Schwartz, S. H. "An Overview of the Schwartz Theory of Basic Values." *Online Readings in Psychology and Culture* 2 (2012). Comprehensive overview of a specific theory of values, understood as motivational orientations toward abstract outcomes (e.g., self-direction, power, security).
- Schwartz, S. H., and K. Boehnke. "Evaluating the Structure of Human Values with Confirmatory Factor Analysis." *Journal of Research in Personality 38* (2004): 230–255. Describes an older method of subjective judgments of relations among valued outcomes and a newer, formal method of analyzing these relations.
- Wallach, W., and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2008. This book describes some of the challenges of having a one-size-fits-all approach to embedding human values in autonomous systems.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

Issue 2:
The need for norm updating.

Background and Analysis

Norms are not static. They change over time, in response to social progress and new legal measures, and, in smaller communities, in response to complaints or new opportunities. New norms form when technological innovation demands novel social standards (e.g., cell phone use in public), and norms can fade away when, for whatever reasons, fewer and fewer people adhere to them.

Humans have many mechanisms available to update norms and learn new ones. They observe other community members' behavior and are sensitive to collective norm change; they explicitly ask about new norms when joining new communities (e.g., entering college, a job in a new town); and they respond to feedback from others when they exhibit uncertainty about norms or have violated a norm.

An A/IS may be equipped with a norm baseline before it is deployed in its target community (Issue 1), but this will not suffice for it to behave appropriately over an extended time. It must be capable of identifying and adding new norms to its baseline system, because the initial norm identification process will undoubtedly have missed some norms. It must also be capable of updating some of its existing norms, as change occurs in its target community. A/IS would be best equipped to respond to such demands for change by relying on multiple mechanisms, such as:

- Processing behavioral trends by members of the target community and comparing them to trends predicted by the baseline norm system;
- Asking for guidance from the community when uncertainty about applicable norms exceeds a critical threshold;
- Responding to instruction from the community members who introduce the robot to a previously unknown context or who notice the A/IS's uncertainty in a familiar context;
- Responding to critique from the community when the A/IS violates a norm.

The modification of a normative system can occur at any level of the system: it could involve altering the priority weightings between individual norms, (changing the qualitative expression of a norm), or altering the quantitative parameters that enable the norm.

As in the case of resolving norm conflicts (Issue 2), we recommend that the system's norm changes be transparent. That is, the system should make explicit when it adds new norms to its norm system or adjusts the priority or content of existing norms. The specific form of communication will vary by machine sophistication (e.g., communication capacity) and function (e.g., flexible social companion vs. task-defined medical robot). In some cases,

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

the system may document its dynamic change and the user can consult this documentation as desired; in other cases, explicit announcements and requests for discussion may be appropriate; in yet other cases, the A/IS may propose changes and the relevant human community will decide whether such changes should be implemented in the system.

Candidate Recommendation

To respond to the dynamic change of norms in society the A/IS must be able to adjust its existing norms and learn new ones, while being transparent about these changes.

Issue 3:

A/IS will face norm conflicts and need methods to resolve them.

Background and Analysis

Often, even within a well-specified context, no action is available that fulfills all obligations and prohibitions. Such situations (often described as moral dilemmas or moral overload; see Van den Hoven, 2012) must be computationally tractable by an A/IS — it cannot simply stop in its tracks and end on a logical contradiction. Humans resolve such situations by accepting trade-offs between conflicting norms, which constitute priorities of one norm or value over another (in a given context). Such priorities may be represented in the norm system as hierarchical relations.

Along with identifying the norms within a specific community and task domain, we need to identify the ways in which people prioritize competing norms and resolve norm conflicts, and the ways in which people expect A/IS to resolve similar norm conflicts. Some general principles are available, such as the Common Good Principle (Andre and Velasquez, 1992). However, other priority relations in the norm network must be established through empirical research so as to reflect the shared values of the community in question. 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.

Some priority orders can be built into a given norm network as hierarchical relations (e.g., prohibitions against harm to humans typically override prohibitions against lying). Other priority orders can stem from the general override that norms in the larger community exert on norms and preferences of an individual user. In the earlier example discussing personalization (see Issue 1), an A/IS of a racist user who demands the A/IS use derogatory language for certain social groups might have to resist such demands because community norms hierarchically override an individual user's preferences.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

In many cases, priority orders are not built in as fixed hierarchies because the priorities are themselves context specific or may arise from net moral costs and benefits of the particular case at hand. A/IS must have learning capacities to track such variations and incorporate user input (e.g., about the subtle differences between contexts) to refine the system's norm network (see Issue 2).

We also recommend that the system's resolution of norm conflicts be transparent — that is, documented by the system and ready to be made available to users. Just like people explain to each other why they made decisions, they will expect any A/IS to be able to explain its decisions (and be sensitive to user feedback about the appropriateness of the decision). To do so, design and development of A/IS should specifically identify the relevant groups of humans who may request explanations and evaluate the system's behavior.

Candidate Recommendation

One must identify the ways in which people resolve norm conflicts and the ways in which they expect A/IS to resolve similar norm conflicts. The system's resolution of norm conflicts must be transparent — that is, documented by the system and ready to be made available to relevant users.

References

- Andre, C., and M. Velasquez. "The Common Good." *Issues in Ethics* 5, no. 1 (1991).
- Van den Hoven, J. "Engineering and the Problem of Moral Overload." *Science and Engineering Ethics* 18, no. 1 (2012): 143–155.

Further Resources

- Abel, D., J. MacGlashan, and M. L. Littman. "Reinforcement Learning as a Framework for Ethical Decision Making." *AAAI Workshop: AI, Ethics, and Society, Volume WS-16-02 of 13th AAAI Workshops*. Palo Alto, CA: AAAI Press, 2016.
- Cushman, F., V. Kumar, and P. Railton. "Moral Learning." *Cognition* 167 (2017): 1–282.
- Open Roboethics Initiative (e.g., on care robots). A series of poll results on differences in human moral decision-making and changes in priority order of values for autonomous systems.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 2 — Implementing Norms in Autonomous Intelligent Systems

Once the norms relevant to an A/IS's role in a specific community have been identified, including their properties and priority structure, we must link these norms to the functionalities of the underlying computational system. We discuss three issues that arise in this process of norm implementation. First, computational approaches to enable a system to represent, learn, and execute norms are only slowly emerging. However, the diversity of approaches may soon lead to substantial advances. Second, for A/IS that operate in human communities, there is a particular need for transparency — ranging from the technical process of implementation to the ethical decisions that A/IS will make in human-machine interactions, which will require a high level of explainability. Third, failures of normative reasoning can be considered inevitable and mitigation strategies should therefore be put in place to handle such failures when they occur. Before we discuss these three issues and corresponding candidate recommendations, we offer one general recommendation for the entire process of implementation:

Candidate Recommendation

Throughout the technical implementation of norms, designers should already consider forms and metrics of evaluation and define and incorporate central criteria for assessing an A/IS's norm conformity (e.g., human-machine agreement on moral decisions, verifiability of A/IS decisions, justified trust).

Issue 1:

Many approaches to norm implementation are currently available, and new ones are being developed.

Background and Analysis

The prospect of developing artificial systems that are sensitive to human norms and factor them into morally or legally significant decisions 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 appears under many names, including: machine morality, machine ethics, moral machines, value alignment, computational ethics, artificial morality, safe AI, and friendly AI.

There are a number of different implementation routes for implementing ethics into autonomous systems. Following Wallach and Allen (2008), we might begin to categorize these as either:

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

A. Top-down approaches, where the system (e.g., a software agent) has some symbolic representation of its activity, and so can identify specific states, plans, or actions as ethical/unethical with respect to particular ethical requirements (e.g., Dennis, Fisher, Slavkovik, Webster, 2016; Pereira and Saptawijaya, 2016; Rötzer, 2016; Scheutz, Malle, and Briggs, 2015); or

B. Bottom-up approaches, where the system (e.g., a learning component) builds up, through experience of what is to be considered ethical/unethical in certain situations, an implicit notion of ethical behavior (e.g., Anderson and Anderson, 2014; Riedl and Harrison, 2016).

Relevant examples of these two are: (A) symbolic agents that have explicit representations of plans, actions, goals, etc.; and (B) machine learning systems that train subsymbolic mechanisms with acceptable ethical behavior. (For more detailed discussion, see Charisi et al., 2017.)

Computers and robots already reflect values in their choices and actions, but these values are programmed or designed in 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. Many of the existing experimental approaches to building moral machines are top-down, in the sense that 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 norms, 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 an AI system cannot presently imitate. Nevertheless, as research on autonomous systems progresses, engineers will explore new ways to either simulate learning capabilities or build alternative mechanisms that fulfill similar functions.

Each of the first two options has obvious limitations, such as option A's inability to learn and adapt and option B's unconstrained learning behavior. A third option tries to address these limitations:

C. Hybrid approaches, combining (A) and (B).

For example, the selection of action might be carried out by a subsymbolic system, but this action must be checked by a symbolic "gateway" agent before being invoked. This is a typical approach for Ethical Governors (Arkin, 2008; Winfield, Blum, and Liu, 2014) or Guardians (Etzioni, 2016) that monitor, restrict, and even adapt certain unacceptable behaviors proposed by the system. (See also Issue 3.) Alternatively, action selection in light of norms could be done in a verifiable logical format, while many of the norms constraining those actions can be learned through bottom-up learning mechanisms (e.g., Arnold, Kasenberg, and Scheutz, 2017).

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

These three architectures are not a comprehensive list of all possible techniques for implementing norms and values into A/IS. For example, some contributors to the multi-agent systems literature have integrated norms into their agent specifications (Andrighetto et al., 2013), and even though these agents live in societal simulations and are too underspecified to be translated into individual A/IS (such as robots), the emerging work can inform cognitive architectures of such A/IS that fully integrate norms. In addition, some experimental approaches may attempt to capture values computationally (Conn, 2017), or attempt to relate norms to values in ways that ground or justify norms (Sommer, 2016). Of course, none of these experimental systems should be deployed outside of the laboratory before testing or before certain criteria are met, which we outline in the remainder of this section and in Section 3.

Candidate Recommendation

In light of the multiple possible approaches to computationally implement norms, diverse research efforts should be pursued, especially collaborative research between scientists from different schools of thought.

Issue 2:

The need for transparency from implementation to deployment.

Background and Analysis

When A/IS are part of social communities and act according to the norms of their communities, people will want to understand the A/IS decisions and actions, just as they want to understand each other's decisions and actions. This is particularly true for morally significant actions or omissions: an ethical reasoning system should be able to explain its own reasoning to a user on request. Thus, transparency (*or explainability*) of A/IS is paramount (Wachter, Mittelstadt, and Floridi, 2017), and it will allow a community to understand, predict, and appropriately trust the A/IS (see Section 1, Issue 2). Moreover, as the norms embedded in A/IS are continuously updated and refined (see Section 1, Issue 2), transparency allows for trust to be maintained (Grodzinsky, Miller, and Wolf 2011), and, where necessary, allows the community to modify a system's norms, reasoning, and behavior.

Transparency can occur at multiple levels (e.g., ordinary language, coder verification) and for multiple stakeholders (e.g., user, engineer, attorney). (See IEEE P7001™, Draft Standard for Transparency of Autonomous Systems.) It should be noted that transparency to all parties may not always be advisable, such as in the case of security programs that prevent a system

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

from being hacked (Kroll et al., 2016). Here we briefly illustrate the broad range of transparency by reference to four ways in which systems can be transparent (traceability, verifiability, nondeception, and intelligibility) and apply these considerations to the implementation of norms in A/IS.

Transparency as traceability. Most relevant for the topic of implementation is the transparency of the software engineering process during implementation (Cleland-Huang, Gotel, and Zisman, 2012). It allows for the originally identified norms (Section 1, Issue 1) to be traced through to the final system. This allows technical inspection of which norms have been implemented, for which contexts, and how norm conflicts are resolved (e.g., priority weights given to different norms). Transparency in the implementation process may also reveal biases that were inadvertently built into systems, such as racism and sexism in search engine algorithms (e.g., Noble, 2013). (See Section 3, Issue 2.) Such traceability in turn calibrates a community's trust about whether A/IS are conforming to the norms and values relevant in its use context (Fleischmann and Wallace, 2005).

Transparency as verifiability. Transparency concerning how normative reasoning is approached in the implementation is important as we wish to verify that the normative decisions the system makes match the required norms and values. Explicit and exact representations of these normative decisions can then provide the basis for a range of strong mathematical techniques, such as formal verification (Fisher, Dennis, and Webster, 2013). Even if a system cannot explain every single reasoning step in understandable human terms, a log of ethical reasoning should be available for inspection of later evaluation purposes.

Transparency as nondeception and honest design. We can assume that lying and deception will be prohibited actions in many contexts, and therefore will be part of the norm system implemented into A/IS. In certain use cases of an A/IS, deception may be necessary in serving the core functionality of the system (e.g., a robot that plays poker with humans), but those actions are no longer norm violations because they are justified by context and user consent.

However, the absence of deception does not yet meet the goal of transparency. One should demand that A/IS be *honest*, and that includes both, more obviously, honest communication by the A/IS itself and, less obviously, "honest design." Honest design entails that the physical appearance of a system accurately represents what the system is capable of doing — e.g., ears only for systems that actually process acoustic information; eyes only for systems that actually process visual information. The requirement for honest design may also extend to higher-level capacities of artificial agents: If the agent introduces a certain topic into conversation, then it should also be able to, if asked, reason about that topic; if the agent displays signs of a certain human-like emotion, then it should have an internal state that corresponds to at least an analogue to that human emotion (e.g., inhabit the appraisal states that make up the emotion).

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

Transparency as intelligibility. As mentioned above, humans will want to understand an A/IS's decisions and actions, especially the morally significant ones. A clear requirement for an ethical A/IS is therefore that the system be able to explain its own reasoning to a user, when asked (or, ideally, also when suspecting the user's confusion), and the system should do so at a level of ordinary human reasoning, not with incomprehensible technical detail (Tintarev and Kutlak, 2014). Furthermore, when the system cannot itself explain some of its actions, technicians or designers should be available to make those actions intelligible. Along these lines, the European Union's new General Data Protection Regulation (GDPR), scheduled to take effect in 2018, states that, for 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." (See Boyd, 2016, for a critical discussion of this regulation.)

Candidate Recommendation

A/IS, and especially those with embedded norms, must have a high level of transparency, from traceability in the implementation process, mathematical verifiability of its reasoning, to honesty in appearance-based signals, and intelligibility of the system's operation and decisions.

Issue 3:
Failures will occur.

Operational failures and, in particular, violations of a system's embedded community norms are unavoidable, both during system testing and during deployment. Not only are implementations never perfect, but A/IS with embedded norms will update or expand their norms over extended use (see Section 1, Issue 2) and interactions in the social world are particularly complex and uncertain. Thus, we propose the following candidate recommendation.

Candidate Recommendation

Because designers cannot anticipate all possible operating conditions and potential failures of A/IS, multiple additional strategies to mitigate the chance and magnitude of harm must be in place.

Elaboration

To be specific, we sample three possible mitigation strategies.

First, anticipating the process of evaluation already during the implementation phase requires defining criteria and metrics for such evaluation, which in turn better allows the detection and mitigation of failures. Metrics will include more technical variables, such as traceability and verifiability; user-level variables such as reliability, understandable explanations, and responsiveness to feedback; and community-level variables such as justified trust (see Issue 2) and the collective

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

belief that A/IS are generally creating social benefits rather than, for example, technological unemployment.

Second, a systematic risk analysis and management approach can be useful (e.g., Oetzel and Spiekermann, 2014, for an application to privacy norms). This approach tries to anticipate potential points of failure (e.g., norm violations) and, where possible, develops some ways to mitigate or remove the effects of failures. Successful behavior, and occasional failures, can then iteratively improve predictions and mitigation attempts.

Third, because not all risks and failures are predictable, especially in complex human-machine interactions in social contexts, additional mitigation mechanisms must be made available. Designers are strongly encouraged to augment the architectures of their systems with components that handle unanticipated norm violations with a fail-safe, such as the symbolic "gateway" agents discussed in Section 1, Issue 1. Designers should identify a number of strict laws (that is, task- and community-specific norms that should never be violated), and the fail-safe components should continuously monitor operations against possible violations of these laws. In case of violations, the higher-order gateway agent should take appropriate actions, such as safely disabling the system's operation until the source of failure is identified. The fail-safe components need to be extremely reliable and protected against security breaches, which can be achieved, for example, by validating them carefully and not letting them adapt their parameters during execution.

References

- Anderson, M., and S. L. Anderson. "GenEth: A General Ethical Dilemma Analyzer." *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014): 253–261.
- Andrighetto, G., G. Governatori, P. Noriega, and L. W. N. van der Torre, eds. *Normative Multi-Agent Systems*. Saarbrücken/Wadern, Germany: Dagstuhl Publishing, 2013.
- Arkin, R. "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/ Reactive Robot Architecture." *Proceedings of the 2008 Conference on Human-Robot Interaction* (2008): 121–128.
- Arnold, T., D. Kasenberg, and M. Scheutz. "Value Alignment or Misalignment — What Will Keep Systems Accountable?" *The Workshops of the Thirty-First AAAI Conference on Artificial Intelligence: Technical Reports, WS-17-02: AI, Ethics, and Society,* 81–88. Palo Alto, CA: The AAAI Press, 2017.
- Boyd, D. "Transparency ≠ Accountability." *Data & Society: Points*, November 29, 2016.
- Charisi, V., L. Dennis, M. Fisher et al. "Towards Moral Autonomous Systems," 2017.
- Cleland-Huang, J., O. Gotel, and A. Zisman, eds. *Software and Systems Traceability.* London: Springer, 2012. doi:10.1007/978-1-4471-2239-5
- Conn, A. "How Do We Align Artificial Intelligence with Human Values?" *Future of Life Institute*, February 3, 2017.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

- Dennis, L., M. Fisher, M. Slavkovik, and M. Webster. "Formal Verification of Ethical Choices in Autonomous Systems." *Robotics and Autonomous Systems* 77 (2016): 1–14.
- Etzioni, A. "Designing AI Systems That Obey Our Laws and Values." *Communications of the ACM* 59, no. 9 (2016): 29–31.
- Fisher, M., L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM* 56, no. 9 (2013): 84–93.
- Fleischmann, K. R., and W. A. Wallace. "A Covenant with Transparency: Opening the Black Box of Models." *Communications of the ACM* 48, no. 5 (2005): 93–97.
- Grodzinsky, F. S., K. W. Miller, and M. J. Wolf. "Developing Artificial Agents Worthy of Trust: Would You Buy a Used Car from This Artificial Agent?" *Ethics and Information Technology* 13, (2011): 17–27.
- Kroll, J. A., J. Huey, J., S. Barocas et al. "Accountable Algorithms." *University of Pennsylvania Law Review* 165 (2017 forthcoming).
- Noble, S. U. "Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible." *InVisible Culture* 19 (2013).
- Oetzel, M. C., and S. Spiekermann. "A Systematic Methodology for Privacy Impact Assessments: A Design Science Approach." *European Journal of Information Systems* 23, (2014): 126–150. doi:10.1057/ejis.2013.18
- Pereira, L. M., and A. Saptawijaya. *Programming Machine Ethics*. Cham, Switzerland: Springer International, 2016.
- Riedl, M. O., and B. Harrison. "Using Stories to Teach Human Values to Artificial Agents." *Proceedings of the 2nd International Workshop on AI, Ethics and Society*, Phoenix, Arizona, 2016.
- Rötzer, F. ed. *Programmierte Ethik: Brauchen Roboter Regeln oder Moral?* Hannover, Germany: Heise Medien, 2016.
- Scheutz, M., B. F. Malle, and G. Briggs. "Towards Morally Sensitive Action Selection for Autonomous Social Robots." *Proceedings of the 24th International Symposium on Robot and Human Interactive Communication, RO-MAN 2015* (2015): 492–497.
- Sommer, A. U. *Werte: Warum Man Sie Braucht, Obwohl es Sie Nicht Gibt*. [Values. Why we need them even though they don't exist.] Stuttgart, Germany: J. B. Metzler, 2016.
- Sommerville, I. *Software Engineering*. Harlow, U.K.: Pearson Studium, 2001.
- Tintarev, N., and R. Kutlak. "Demo: Making Plans Scrutable with Argumentation and Natural Language Generation." *Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces* (2014): 29–32.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

- Wachter, S., B. Mittelstadt, and L. Floridi. "Transparent, Explainable, and Accountable AI for Robotics." *Science Robotics* 2, no. 6 (2017): eaan6080. doi:10.1126/scirobotics.aan6080
- Wallach, W., and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2008.
- Winfield A. F. T., C. Blum, and W. Liu. "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection" in *Advances in Autonomous Robotics Systems, Lecture Notes in Computer Science Volume*, edited by M. Mistry, A. Leonardis, M. Witkowski, and C. Melhuish, 85–96. Springer, 2014.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 3 — Evaluating the Implementation of A/IS

The success of implementing appropriate norms in A/IS must be rigorously evaluated. This evaluation process must be anticipated during design and incorporated into the implementation process, and it must continue throughout the life cycle of the system's deployment. Assessment before full-scale deployment would best take place in systematic test beds that allow human users (from the defined community, and representing all demographic groups) to engage safely with the A/IS in intended tasks. Multiple disciplines and methods should contribute to developing and conducting such evaluations.

Evaluation criteria must capture the quality of human-machine interactions, human approval and appreciation of the A/IS, trust in the A/IS, adaptability of the A/IS to human users, and human benefits in the presence or under the influence of the A/IS. A range of ethical/normative aspects to be considered can be found in the UK standard on Robot Ethics (BSI, 2016). These are important general evaluation criteria, but they do not yet fully capture evaluation of a system that has *norm capacities*. To evaluate a system's norm-conforming behavior, one must describe (and ideally, formally specify) criterion behaviors that reflect the previously identified norms, describe what the user expects the system to do, verify that the system really does this, and validate that the specification actually matches the criteria. Many different evaluation techniques are available in the field of software engineering (Sommerville, 2001), ranging from formal mathematical proof, through rigorous empirical testing against criteria of normatively correct behavior, to informal analysis of user interactions and responses to the machine's norm awareness and compliance. All these approaches can, in principle, be applied to the full range of autonomous systems, including robots (Fisher, Dennis, and Webster, 2013).

Evaluation may be done by first parties (designers/manufacturers, and users) as well as third parties (e.g., regulators or independent testing agencies). In either case, the results of evaluations should be made available to all parties, with strong encouragement to resolve discovered system limitations and resolve potential discrepancies among multiple evaluations.

Candidate Recommendation

Evaluation must be anticipated during a system's design, incorporated into the implementation process, and continue throughout the system's deployment. Evaluation must include multiple methods, be made available to all parties (from designers and users to regulators), and should include procedures to resolve conflicting evaluation results.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

Issue 1:

Not all norms of a target community apply equally to human and artificial agents.

Background and Analysis

An intuitive criterion for evaluations of norms embedded in A/IS would be that the A/IS norms should mirror the community's norms — that is, the A/IS should be disposed to behave the same way that people expect each other to behave. However, for a given community and a given A/IS use context, A/IS and humans may not have *identical* sets of norms. People will have some unique expectations for humans than they do for machines (e.g., norms governing the regulation of negative emotions, assuming that machines do not have such emotions), and people will have some unique expectations of A/IS that they do not have for humans (e.g., that the machine will sacrifice itself, if it can, to prevent harm to a human).

Candidate Recommendation

The norm identification process must document the similarities and differences between the norms that humans apply to other humans and the norms they apply to A/IS. Norm implementations should be evaluated specifically against the norms that the community expects the A/IS to follow.

Issue 2:

A/IS can have biases that disadvantage specific groups.

Background and Analysis

Even when reflecting the full system of community norms that was identified, A/IS may show operation biases that disadvantage specific groups in the community or instill biases in users by reinforcing group stereotypes. A system's bias can emerge in perception (e.g., a passport application AI rejected an Asian man's photo because it insisted his eyes were closed; Griffiths, 2016); information processing (e.g., speech recognition systems are notoriously less accurate for female speakers than for male speakers; Tatman, 2016); decisions (e.g., a criminal risk assessment device overpredicts recidivism by African Americans; Angwin, et al., 2016); and even in its own appearance and presentation (e.g., the vast majority of humanoid robots have white "skin" color and use female voices) (Riek and Howard, 2014).

The norm identification process detailed in Section 1 is intended to minimize individual designers' biases, because the community norms are assessed empirically. The process also seeks to incorporate values and norms against prejudice and discrimination. However, biases may still emerge from imperfections in the norm identification process itself, from unrepresentative training sets for machine learning systems, and from programmers' and designers' unconscious

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

assumptions. Therefore, unanticipated or undetected biases should be further reduced by including members of diverse social groups in both the planning and evaluation of AI systems and integrating community outreach into the evaluation process (e.g., DO-IT program; RRI framework). Behavioral scientists and members of the target populations will be particularly valuable when devising criterion tasks for system evaluation. Such tasks would assess, for example, whether the A/IS applies norms in discriminatory ways to different races, ethnicities, genders, ages, body shapes, or to people who use wheelchairs or prosthetics, and so on.

Candidate Recommendation

Evaluation of A/IS must carefully assess potential biases in the system's performance that disadvantage specific social groups. The evaluation process should integrate members of potentially disadvantaged groups to diagnose and correct such biases.

Issue 3:
Challenges to evaluation by third parties.

Background and Analysis

A/IS should have sufficient transparency to allow evaluation by third parties, including regulators, consumer advocates, ethicists, post-accident investigators, or society at large. However, transparency can be severely limited in some systems, especially in those that rely on machine learning algorithms trained on large data sets. The data sets may not be accessible to evaluators; the algorithms may be proprietary information or mathematically so complex that they defy common-sense explanation; and even fellow software experts may be unable to verify reliability and efficacy of the final system because the system's specifications are opaque.

For less inscrutable systems, numerous techniques are available to evaluate the implementation of an A/IS's norm conformity. On one side there is formal verification, which provides a mathematical proof that the A/IS will always match specific normative and ethical requirements (typically devised in a top-down approach; see Section 2, Issue 1). This approach requires access to the decision-making process and the reasons for each decision (Fisher, Dennis, and Webster, 2013). A simpler alternative, sometimes suitable even for machine learning systems, is to test the A/IS against a set of scenarios and assess how well it matches its normative requirements (e.g., acting in accordance with relevant norms; recognizing other agents' norm violations).

These different evaluation techniques can be assigned different levels of "strength" — strong ones demonstrate the exhaustive set of an A/IS's allowable behaviors for a range of criterion scenarios; weaker ones sample from criterion scenarios and illustrate the system's behavior for that subsample. In the latter case, confidence in the A/IS's ability to meet normative requirements is more limited. An evaluation's

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

concluding judgment must therefore acknowledge the strength of the verification technique used, and the expressed confidence in the evaluation (and in the A/IS itself) must be qualified by this level of strength.

Transparency is only a necessary requirement for a more important long-term goal, having systems be accountable to their users and community members. However, this goal raises many questions such as to whom the A/IS are accountable and who has the right to correct the systems, or also which kind of A/IS should be subject to accountability requirements.

Candidate Recommendation

To maximize effective evaluation by third parties (e.g., regulators, accident investigators), A/IS should be designed, specified, and documented so as to permit the use of strong verification and validation techniques for assessing the system's safety and norm compliance, in order to possibly achieve accountability to the relevant communities.

References

- Angwin, J., J. Larson, S. Mattu, L. Kirchner. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks." *ProPublica*, May 23, 2016.
- British Standards Institution. BS8611:2016, "Robots and Robotic Devices. Guide to the Ethical Design and Application of Robots and Robotic Systems," 2016.
- Federal Trade Commission. "Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues. FTC Report." Washington DC: Federal Trade Commission, 2016.
- Fisher, M., L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM* 56 (2013): 84–93.
- Griffiths, J. "New Zealand Passport Robot Thinks This Asian Man's Eyes Are Closed." *CNN.com*, December 9, 2016.
- Riek, L. D., and D. Howard. "A Code of Ethics for the Human-Robot Interaction Profession." *Proceedings of We Robot*, April 4, 2014.
- Tatman, R. "Google's Speech Recognition Has a Gender Bias." *Making Noise and Hearing Things*, July 12, 2016.

Further Resources

- Anderson, M., and S. L. Anderson eds. Machine Ethics. New York: Cambridge University Press, 2011.
- Abney, K., G. A. Bekey, and P. Lin. Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge, MA: The MIT Press, 2011.
- Boden, M., J. Bryson et al. "Principles of Robotics: Regulating Robots in the Real World." *Connection Science* 29, no. 2 (2017): 124–129.
- Coeckelbergh, M. "Can We Trust Robots?" *Ethics and Information Technology* 14 (2012): 53–60.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Embedding Values into Autonomous Intelligent Systems

- Dennis, L. A., M. Fisher, N. Lincoln, A. Lisitsa, and S. M. Veres. "Practical Verification of Decision-Making in Agent-Based Autonomous Systems." *Automated Software Engineering* 23, no. 3, (2016): 305–359.
- Fisher, M., C. List, M. Slavkovik, and A. F. T. Winfield. "Engineering Moral Agents — From Human Morality to Artificial Morality" (Dagstuhl Seminar 16222). *Dagstuhl Reports* 6, no. 5 (2016): 114–137.
- Fleischmann, K. R. *Information and Human Values*. San Rafael, CA: Morgan and Claypool, 2014.
- Governatori, G., and A. Rotolo. "How Do Agents Comply with Norms?" in *Normative Multi-Agent Systems*, edited by G. Boella, P. Noriega, G. Pigozzi, and H. Verhagen, Dagstuhl Seminar Proceedings. Dagstuhl, Germany: Schloss Dagstuhl — Leibniz-Zentrum fuer Informatik, 2009.
- Leet, E. H., and W. A. Wallace. "Society's Role and the *Ethics of Modeling*," in Ethics in Modeling, edited by W. A. Wallace, 242–245. Tarrytown, NY: Elsevier, 1994.
- Jarvenpaa, S. L., N. Tractinsky, and L. Saarinen. "Consumer Trust in an Internet Store: A Cross-Cultural Validation." *Journal of Computer-Mediated Communication* 5, no. 2 (1999): 1–37.
- Mahmoud, M. A., M. S. Ahmad, M. Z. Mohd Yusoff, and A. Mustapha. "A Review of Norms and Normative Multiagent Systems." *The Scientific World Journal*, (2014): 1–23.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

To ensure autonomous and intelligent systems (A/IS) are aligned to benefit humanity A/IS research and design must be underpinned by ethical and legal norms as well as methods. We strongly believe that a value-based design methodology should become the essential focus for the modern A/IS organization.

Value-based system design methods put human advancement at the core of A/IS development. Such methods recognize that machines should serve humans, and not the other way around. A/IS developers should employ value-based design methods to create sustainable systems that are thoroughly scrutinized for social costs and advantages that will also increase economic value for organizations. To create A/IS that enhances human well-being and freedom, system design methodologies should also be enriched by putting greater emphasis on internationally recognized human rights, as a primary form of human values.

To help achieve these goals, researchers and technologists need to embrace transparency regarding their processes, products, values, and design practices to increase end-user and community trust. It will be essential that educational institutions inform engineering students about ethics, justice, and human rights, address ethical research and business practices surrounding the development of A/IS, and attend to the responsibility of the technology sector vis-à-vis public interest issues. The proliferation of value-based design will require a change of current system development approaches for organizations, including a commitment of research institutions to strong ethical guidelines for research, and of businesses to values that transcend narrow economic incentives.

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 1 — Interdisciplinary Education and Research

Integrating applied ethics into education and research to address the issues of autonomous and intelligent systems (A/IS) requires an interdisciplinary approach, bringing together humanities, social sciences, science, engineering, and other disciplines.

Issue:

Inadequate integration of ethics in A/IS-related degree programs.

Background

AI engineers and design teams too often fail to thoroughly explore the ethical considerations implicit in their technical work and design choices. They tend to treat ethical decision-making as another form of technical problem solving. Although ethical challenges often have technical solutions, identifying and ameliorating those challenges requires technicians to methodically inquire about the social context of their work. 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 and simplified in the context of technical practice. This originates in the fact that many engineering programs do not sufficiently integrate coursework, training, or practical experience in applied ethics throughout their curricula; too often ethics is relegated to a stand-alone course or module that gives students little or no direct experience in ethical decision-making in engineering work. Ethics education for engineering students should be meaningful, measurable, and incorporate best practices of STEM ethics education drawn from pertinent multidisciplinary resources.

The aim of these recommendations is to prepare students for the technical training and engineering development methodologies that incorporate ethics as essential so that ethics and human rights become naturally part of the design process.

Candidate Recommendations

Ethics and ethical reflection need to be a core subject for aspiring engineers and technologists beginning at the earliest appropriate level and for all advanced degrees. By training students how to be sensitive to ethical issues in design before they enter the workplace, they can more effectively implement value-based design methodologies in the context of A/IS work.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

We also recommend that effective STEM ethics curricula be informed by *scientists, artists, philosophers, psychologists, legal scholars, engineers, and other subject matter experts* from a variety of cultural backgrounds to ensure that students acquire sensitivity to a diversity of robust perspectives on human flourishing. Such curricula should teach aspiring engineers, computer scientists, and statisticians about the relevance and impact of their decisions in designing A/IS technologies. Effective ethics education in STEM contexts should span primary, secondary, and post-secondary education, and include both universities and vocational training schools. Relevant accreditation bodies should reinforce this integrated approach as outlined above.

Further Resources

- Holdren, J., and M. Smith. "Preparing for the Future of Artificial Intelligence." Washington, DC: Executive Office of the President, National Science and Technology Council, 2016. This White House report 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 programs.
- The U.S. National Science Foundation has funded extensive research on STEM ethics education best practices through the Cultivating Cultures for Ethical Science, Technology, Engineering, and Mathematics (CCE-STEM) Program, and recommends integrative approaches that incorporate ethics throughout STEM education.
- Comparing the UK, EU, and US approaches to AI and ethics: Cath, C. et al. "Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach." *Science and Engineering Ethics* (2017).
- The Oxford Internet Institute (OII) organized a workshop on ethical issues in engineering. The output paper can be found here: Zevenbergen, B. et al. "Philosophy Meets Internet Engineering: Ethics in Networked Systems Research." Oxford, U.K.: Oxford Internet Institute, University of Oxford, 2015.
- Companies should also be encouraged to mandate consideration of ethics at the pre-product design stage, as was done by Lucid AI.
- There are a variety of peer-reviewed online resources collecting STEM ethics curricula, syllabi, and education modules:
  - Ethics Education Library, Illinois Institute of Technology
  - IDEESE: International Dimensions of Ethics Education in Science & Engineering, University of Massachusetts Amherst
  - National Center for Professional & Research Ethics, University of Illinois
  - Online Ethics Center, National Academy of Engineering

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

Issue:

The need for more constructive and sustained interdisciplinary collaborations to address ethical issues concerning autonomous and intelligent systems (A/IS).

Background

Not enough institutional resources and incentive structures exist for bringing A/IS engineers and designers into sustained and constructive contact with ethicists, legal scholars, and social scientists, both in academia and industry. This contact is necessary as it can enable meaningful interdisciplinary collaboration to shape the future of technological innovation. There are currently few methodologies, shared knowledge, and lexicons that would facilitate such collaborations.

This issue, to a large degree, relates to funding models as well as the traditional mono-function culture in A/IS-related institutions and companies, which limit cross-pollination between disciplines (see below). To help bridge this gap, additional "translation work" and resource sharing (including websites and MOOCs) needs to happen among technologists and other relevant experts (e.g., in medicine, architecture, law, philosophy, psychology, cognitive science).

Candidate Recommendations

Funding models and institutional incentive structures should be reviewed and revised to prioritize projects with interdisciplinary ethics components to encourage integration of ethics into projects at all levels.

Further Resources

- Barocas, S. Course Material for Ethics and Policy in Data Science.
- Floridi, L., and M. Taddeo. "What Is Data Ethics?" *Philosophical Transactions of the Royal Society* 374, no. 2083 (2014): 1–4. doi:10.1098/rsta.2016.0360.
- Spiekermann, S. *Ethical IT Innovation: A Value-Based System Design Approach*. Boca Raton, Florida: Auerbach Publications, 2015.
- The approach developed by the Internet Research Task Force's Human Rights Protocol Research Group (HRPC) for integrating human rights concern in technical design.

Issue:

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

Background

A responsible approach to embedded values (both as uncritical bias and as value by design) in information and communications technology (ICT), algorithms and autonomous systems will need to differentiate between culturally distinctive values (i.e., how do different cultures

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

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 oversimplified ethical relativism, or embedding values that are antithetical to human flourishing (for example, human rights violations), it is critical that A/IS design avoids only considering monocultural influenced ethical foundations.

Candidate Recommendations

Establish a leading role for intercultural information ethics (IIE) practitioners in 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.

Further Resources

- Pauleen, D. 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, no. 17 (2006). The work of Pauleen et al. (2006) and Bielby (2015) has been guiding in this field: "Cultural values, attitudes, and behaviours prominently influence how a given group of people views, understands, processes, communicates, and manages data, information, and knowledge."
- Bielby, J. "Comparative Philosophies in Intercultural Information Ethics," *Confluence: Online Journal of World Philosophies* 2, no. 1 (2015): 233–253.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 2 – Corporate Practices and A/IS

Corporations, whether for-profit or not-for-profit, are eager to develop, deploy, and monetize A/IS, but there are insufficient structures in place for creating and supporting ethical systems and practices around A/IS funding, 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 and intelligent systems (A/IS). To do so, we need to further identify and refine social processes and management strategies that facilitate values-based design in the engineering and manufacturing process.

Candidate Recommendations

The building blocks of such practices include top-down leadership, bottom-up empowerment, ownership, and responsibility, and the need to consider system deployment contexts and/or ecosystems. The institution of an ethical A/IS corporate culture would accelerate the adoption of the other recommendations 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.
- *Firms of Endearment* is a book which showcases how companies embracing values and a stakeholder approach outperform their competitors in the long run.
- The ACM Code of Ethics and Professional Ethics, which also includes various references to human well-being and human rights.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

Issue:

Lack of values-aware leadership.

Background

Technology leadership should give innovation teams and engineers direction regarding which human values and legal norms should be promoted in the design of an A/IS system. Cultivating an ethical corporate culture is an essential component of successful leadership in the A/IS domain.

Candidate Recommendations

Companies need to create roles for senior-level marketers, ethicists, or lawyers who can pragmatically implement ethically aligned design, both the technology and the social processes to support value-based system innovation. Companies need to ensure that their understanding of value-based system innovation is based on *de jure* and *de facto* international human rights standards.

A promising way to ensure values are on the agenda in system development is to have a Chief Values Officer (CVO), a role first suggested by Kay Firth-Butterfield, Vice-Chair, The IEEE Global Initiative and Project Head of AI and Machine Learning at the World Economic Forum. The CVO should support system innovations and engineering teams to consider values and provide them with methodological guidance on how to do so. However, ethical responsibility should not be delegated solely to CVOs. 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.

Further Resources

- United Nations, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, New York and Geneva: UN, 2011.
- Institute for Human Rights and Business (IHRB), and Shift, SectICTor Guide on Implementing the UN Guiding Principles on Business and Human Rights, 2013.
- Cath, C., and L. Floridi. "The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights." *Science and Engineering Ethics* 23, no. 2 (2017): 449–468.
- Butterfield, Kay-Firth (2017). How IEEE Aims to Instill Ethics in Artificial Intelligence Design. *The Institute*.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

Issue:

Lack of empowerment to raise ethical concerns.

Background

Engineers and design teams can encounter obstacles to raising ethical concerns regarding their designs or design specifications within their organizations. Corporate culture should incentivize technical staff to voice the full range of ethical questions to relevant corporate actors throughout the full product lifecycle. Because raising ethical concerns can be perceived as slowing or halting a design project, organizations need to consider how they can recognize and incentivize value-based design as an integral component of product development.

Candidate Recommendations

Employees should be empowered to raise ethical concerns in day-to-day professional practice, not just in extreme emergency circumstances such as whistleblowing. New organizational and socio-cultural processes that broaden the scope around professional ethics and design need to be implemented within organizations. New categories of considerations around these issues need to be accommodated along with new forms of Codes of Conduct, so individuals are empowered to share their insights and concerns in an atmosphere of trust.

Further Resources

- The British Computer Society (BCS) code of conduct holds that individuals have to: "a) have due regard for public health, privacy, security and well-being 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, colour, 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."
- 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 as they relate to human rights 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.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

Issue:

Organizations should examine their cultures to determine how to flexibly implement value-based design.

Background

Ethics is often treated as an impediment to innovation, even among those who ostensibly support ethical design practices. In industries that reward rapid innovation, it is necessary to develop design practices that integrate effectively with existing engineering workflows. Those who advocate for ethical design within a company should not be seen as innovators seeking the best ultimate outcomes for the company, end-users, and society. Leaders can facilitate that mindset by promoting an organizational structure that supports the integration of dialogue about ethics throughout product lifecycles.

A/IS design processes often present moments where ethical consequences can be highlighted. There are no universally prescribed models for this because organizations vary significantly in structure and culture. In some organizations, design team meetings may be brief and informal. In others, the meetings may be lengthy and structured. Regardless, team members should understand how to raise such questions without being perceived as impediments by peers and managers. The transitions point between discovery, prototyping, release, and revisions are natural contexts for conducting such reviews.

Iterative review processes are also advisable, in part because changes to risk profiles over time can illustrate needs or opportunities for improving the final product.

Candidate Recommendations

Companies should study their own design processes to identify moments where engineers and researchers can be encouraged to raise and resolve questions of ethics. Achieving a distributed responsibility for ethics requires that all people involved in product design are encouraged to notice and respond to ethical concerns, particularly around safety, bias, and legality. Organizations should consider how they can best encourage and accommodate lightweight deliberations among peers.

Additionally, organizations should identify points for formal review inside their product development processes. These reviews can focus on "red flags" that have been identified in advance as indicators of risk. For example, if the datasets involve minors or focus on users from protected classes then it may require additional justification or alterations to the research or development protocols.

Further Resources

- Sinclair, A. "Approaches to Organizational Culture and Ethics." *Journal of Business Ethics* 12, no. 1 (1993): 63–73.
- Chen, A. Y. S., R. B. Sawyers, and P. F. Williams. "Reinforcing Ethical Decision Making Through Corporate Culture." *Journal of Business Ethics* 16, no. 8 (1997): 855–865.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

- Crawford, K., and R. Calo. "There Is a Blind Spot in AI Research." *Nature* 538 (2016): 311–313.

Issue:

Lack of ownership or responsibility from the tech community.

Background

There is a divergence between the values the technology community sees as its responsibility in regards to A/IS, and the broader set of social concerns raised by the public, legal, and professional communities. The current makeup of most organizations has clear delineations among engineering, legal, and marketing arenas. Thus technologists feel responsible for safety issues regarding their work, but for larger social issues may say, "legal will handle that." In addition, in employment and management technology or work contexts, "ethics" typically refers to a code of conduct regarding professional decorum (versus a values-driven design process mentality). As such, ethics regarding professional conduct often implies moral issues such as integrity or the lack thereof (in the case of whistleblowing, for instance), but ethics in A/IS design includes broader considerations about the consequences of technologies.

Candidate Recommendations

To help integrate applied ethics regarding A/IS and in general, organizations need to choose specific language that will break down traditional biases or barriers and increase adoption of values-based design. For instance, an organization can refer to the "trade-offs" (or "value trade-offs") involved in the examination of the fairness of an algorithm to a specific end user population.

Organizations should clarify the relationship between professional ethics and applied A/IS ethics and help designers, engineers, and other company representatives discern the differences between them and where they complement each other.

Corporate ethical review boards, or comparable mechanisms, should be formed to address ethical concerns in relation to their A/IS research. Such boards should seek an appropriately diverse composition and use relevant criteria, including both research ethics and product ethics at the appropriate levels of advancement of research and development. These boards should examine justifications of research or industrial projects in terms of consequences for human flourishing.

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 approach to the implementation of ethics within an organization and describes Facebook's internal ethics review for research and development.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

- 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 drawn can also be applied to ethical review boards. Examples of organizations dealing with such trade-offs can for instance be found in the security considerations of the Internet Engineering Task Force (IETF).

Issue:

Need to include stakeholders for adequate ethical perspective on A/IS.

Background

The interface between AI and practitioners, as well as other stakeholders, is gaining broader attention in domains such as health care diagnostics, and there are many other contexts 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. Technologists need to have that stakeholder feedback, because beyond academically oriented language about ethics, that feedback is often about crucial design detail gained by experience (form, sound, space, dialogue concepts). There are successful models of user experience (UX design) that account for human factors which should be incorporated to A/IS design as systems are more widely deployed.

Candidate Recommendations

Account for the interests of the full range of stakeholders or practitioners who will be working alongside A/IS, incorporating their insights. Build upon, rather than circumvent or ignore, the social and practical wisdom of involved practitioners and other stakeholders.

Further Resources

- Schroeter, Ch. et al. "Realization and User Evaluation of a Companion Robot for People with Mild Cognitive Impairments." *Proceedings of IEEE International Conference on Robotics and Automation (ICRA 2013)*, Karlsruhe, Germany (2013): 1145–1151.
- Chen, T. L. et al. "Robots for Humanity: Using Assistive Robotics to Empower People with Disabilities." *IEEE Robotics and Automation Magazine* 20, no. 1 (2013): 30–39.
- Hartson, R., and P. S. Pyla. *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Waltham, MA: Elsevier, 2012.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 3 — Research Ethics for Development and Testing of A/IS Technologies

Issue:

Institutional ethics committees are under-resourced to address the ethics of R&D in the A/IS fields.

Background

It is unclear how research on the interface of humans and A/IS, animals and A/IS, and biological hazards will pose practical challenges for research ethical review boards. Norms, institutional controls, and risk metrics appropriate to the technology are not well established in the relevant literature and research governance infrastructure. Additionally, national and international regulations governing review of human-subjects research may explicitly or implicitly exclude A/IS research from their purview on the basis of legal technicalities or medical ethical concerns regardless of potential harms posed by the research.

Research on A/IS human-machine interaction, when it involves intervention or interaction with identifiable human participants or their data, typically falls to the governance of research ethics boards (e.g., institutional review boards). The national level and institutional resources (e.g., hospitals and universities) to govern ethical conduct of HCI, particularly within the disciplines pertinent to A/IS research, are underdeveloped. First, there is limited international or national guidance to govern this form of research. While sections of IEEE standards governing research on AI in medical devices address some of the issues related to security of AI-enabled devices, the ethics of testing those devices to bring them to market are not developed into recognized national (e.g., U.S. FDA) or international (e.g., EU EMA) policies or guidance documents. Second, the bodies that typically train individuals to be gatekeepers for the research ethics bodies (e.g., PRIM&R, SoCRA) are under-resourced in terms of expertise for A/IS development. Third, it is not clear whether there is sufficient attention paid to A/IS ethics by research ethics board members or by researchers whose projects involve the use of human participants or their identifiable data.

Research pertinent to the ethics governing research at the interface of animals and A/IS research is underdeveloped with respect to systematization for implementation by

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

IACUC or other relevant committees. In institutions without a veterinary school, it is unclear that the organization would have the relevant resources necessary to conduct an ethical review of such research.

Research pertinent to the intersection of radiological, biological, and toxicological research (ordinarily governed under institutional biosafety committees) and A/IS research is not found often in the literature pertinent to research ethics or research governance. Beyond a limited number of pieces addressing the "dual use" or import/export requirements for A/IS in weapons development, there are no guidelines or standards governing topics ordinarily reserved for review by institutional biosafety committees, or institutional radiological safety committees, or laboratory safety committees.

Candidate Recommendations

IEEE should draw upon existing standards, empirical research, and expertise to identify priorities and develop standards for governance of A/IS research and to partner with relevant national agencies, and international organizations, when possible.

Further Resources

- Jordan, S. R. "The Innovation Imperative." *Public Management Review* 16, no. 1 (2014): 67–89.
- Schneiderman, B. "The Dangers of Faulty, Biased, or Malicious Algorithms Requires Independent Oversight." *Proceedings of the National Academy of Sciences of the United States of America* 113, no. 48 (2016): 13538–13540.
- Metcalf, J., and K. Crawford. "Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide." SSRN Scholarly Paper, Rochester, NY: Social Science Research Network, 2016.
- Calo, R. "Consumer Subject Review Boards: A Thought Experiment." *Stanford Law Review Online* 66 (2013): 97.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 4 — Lack of Transparency

Lack of transparency about the A/IS 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 Recommendation

Software engineers should be required to document all of their systems and related data flows, their performance, 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, and readability.

Further Resources

- Cath, C. J. N., L. Glorioso, and M. R. Taddeo. "NATO CCD COE Workshop on 'Ethics and Policies for Cyber Warfare'" NATO Cybersecurity Centre for Excellence (CCDCOE) Report. Oxford, U.K.: Magdalen College. Addressed indicators of transparency along these lines.
- Turilli, M., and L. Floridi. "The Ethics of Information Transparency." *Ethics and Information Technology* 11, no. 2 (2009): 105–112.
- Wachter, S., B. Mittelstadt, and L. Floridi. "Transparent, Explainable, and Accountable AI for Robotics." *Science Robotics* 2, no. 6 (2017).
- Kroll, J. A., J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. "Accountable Algorithms." *University of Pennsylvania Law Review* 165, no. 1 (2017): 633–705.
- Balkin, J. M., "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation." *UC Davis Law Review*, (2018 forthcoming).

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

Issue:

Inconsistent or lacking oversight for algorithms.

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. These recommendations are similar to those made in committees 1 and 2, but here are used as they apply to the narrow scope of this group.

Candidate Recommendations

Accountability

As touched on in the General Principles section of Ethically Aligned Design, algorithmic 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 harm and negative consequences of the use of these technologies. Here we can look to other technical domains, such as biomedical, civil, and aerospace engineering, where commercial protections for proprietary technology are routinely and effectively balanced with the need for appropriate oversight standards and mechanisms to safeguard the public.

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.
- Ryan Calo, Associate Professor of Law at the University of Washington, wrote an excellent article that gives a detailed overview of a broad array of AI policy questions.
- In the United States, a recent court case, Armstrong, highlights the need for appropriate oversight of algorithmic decision-making, to preserve due process and other legal and ethical principles. K.W. v. Armstrong, 180 F. Supp. 3d 703 (D. Idaho 2016). In the case, a court ruled that Idaho's Department of Health and Welfare violated the rights of disabled Medicaid recipients by relying upon arbitrary and flawed algorithmic decision systems when cutting benefits, and refusing to disclose the decision bases as 'trade secrets.' See details of the case here: https://www.aclu.org/news/federal-court-rules-against-idaho-department-health-and-welfare-medicaid-class-action and a related discussion of the general risks of opaque algorithmic bureaucracies here: https://medium.com/aclu/pitfalls-of-artificial-intelligence-decisionmaking-highlighted-in-idaho-aclu-case-ec59941fb026

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

Issue:
Lack of an independent review organization.

Background

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

For example, today it is possible for systems to download new self-parking functionality 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. We also need further government funding for research into how A/IS technologies can best be subjected to review, and how review organizations can consider both traditional health and safety issues, as well as ethical considerations.

Candidate Recommendations

An independent, internationally coordinated body should be formed to oversee whether such products actually meet ethical criteria, both when deployed, and considering their evolution after deployment and interaction with other products.

Further Resources

- Tutt, A. "An FDA for Algorithms." *Administrative Law Review* (2017): 83–123.
- Scherer, M. U. "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." *Harvard Journal of Law and Technology* 29, no. 2 (2016): 354–400.
- Desai, D. R., and J. A. Kroll. "Trust But Verify: A Guide to Algorithms and the Law." *Harvard Journal of Law and Technology* (2018 forthcoming).

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"

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

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 Recommendation

When systems are built that could impact the safety or well-being 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.

Candidate Recommendation

Technologists should be able to characterize what their algorithms or systems are going to do via transparent and traceable standards. To the degree possible, these characterizations should be predictive, but given the nature of A/IS, they might need to be more retrospective and mitigation oriented. Such standards may include preferential adoption of effective design methodologies for building "explainable AI" (XAI) systems that can provide justifying reasons or other reliable "explanatory" data illuminating the cognitive processes leading to, and/or salient bases for, their conclusions.

Candidate Recommendation

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

Candidate Recommendation

Software engineers should employ "black-box" (opaque) 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*. Cambridge, MA: Harvard University Press, 2015.
- In the United States, in addition to similar commercial endeavors by Oracle and other companies, DARPA (Defense Advanced Research Projects Agency) recently funded a 5-year research program in explainable AI (XAI) methodologies.
- Ananny, M., and K. Crawford. (2016). "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society*, December 13, 2016.
- Another excellent resource on these issues can be found in Chava Gourarie's "Investigating the Algorithms That Govern Our Lives." *Columbia Journalism Review*, April 14, 2016. These recommended reads come at the end of the article:

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Methodologies to Guide Ethical Research and Design

- "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 gentle guide to thinking about data as communication, much of which applies to reporting on algorithms as well.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

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

The concept of intelligence can be difficult to precisely define, and there are many proposed definitions. Legg and Hutter (2007) surveyed 70-odd definitions of intelligence, pulling out the key features and commonalities between them, and settled on the following: "intelligence measures an agent's ability to achieve goals in a wide range of environments."

In the context of autonomous and intelligent systems (A/IS), artificial general intelligence (AGI) is often used to refer to A/IS that perform comparably to humans on intellectual tasks, and artificial superintelligence (ASI or superintelligence) is commonly defined as "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" (Bostrom 2014), passing some threshold of generality, well-roundedness, and versatility that present-day AI systems do not yet achieve.

Although today's state-of-the-art A/IS do not match humans in this capacity (since today's systems are only capable of performing well in limited and narrow environments or domains), many independent researchers and organizations are working on creating AGI systems (including leading AI labs like DeepMind, OpenAI, Microsoft, and Facebook's FAIR), and most AI experts expect A/IS to surpass human-level intelligence sometime this century (Grace et al. 2017).

When reasoning about the impacts that AGI systems will have, it is tempting to anthropomorphize, assume that these systems will have a "mind" similar to that of a human, and conflate intelligence with consciousness. Although it should be possible to build AGI systems that imitate the human brain, the human brain represents one point in a vast space of possible minds (Yampolskiy 2015). AGI systems will not be subject to the same constraints and engineering trade-offs as the human brain (a product of natural selection). Thus, we should not expect AGI systems to necessarily resemble human brains, just as we don't expect planes to resemble birds, even though both are flying machines. This also means that familiar faculties of intelligent entities we know like morality, compassion, and common sense will not be present by default in these new intelligences.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

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

History shows that the largest drivers of change in human welfare, for better and for worse, have been developments in science, technology, and economics. Humanity's ability to drive this change is largely a function of our intelligence. Thus, one can think about building AGI as automating scientific, technological, and economic innovation. Given the disproportionate impact our intelligence has enabled our species to have on the planet and our way of life, we should expect AGI systems to have a disproportionate impact on our future, on a scale not seen since the Industrial Revolution. As such, the development of AGI systems and improvements of those systems toward superintelligence could bring about unprecedented levels of global prosperity. However, it is by no means guaranteed that the impact of these systems will be a positive one without a concerted effort by the A/IS community and other key stakeholders to align them with our interests.

As with other powerful technologies, the development and use of A/IS have always involved risk, either because of misuse or poor design (as simple examples being an assembly line worker being injured by a robotic arm or a guard robot running over a child's foot). However, as systems approach and surpass AGI, unanticipated or unintended system behavior (due to, e.g., architecture choices, training or goal specification failures, mistakes in implementation, or mistaken assumptions) will become increasingly dangerous and difficult to correct. It is likely that not all AGI-level A/IS architectures are alignable with human interests, and as such, care should be taken to analyze how different architectures will perform as they become more capable. In addition to these technical challenges, technologists will also confront a progressively more complex set of ethical issues during the development and deployment of these technologies.

In section 1 which focuses on technical issues, we recommend that A/IS 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 researchers and developers, and to evaluate relevant projects and their progress. In Section 2 which focuses on general principles, we recommend that the A/IS community encourage and promote the sharing

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

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

of safety-related research and tools, and that all those involved in the development and deployment take on the norm that future highly capable transformative A/IS "should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization." (Future of Life Institute 2017)

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 1 — Technical

Issue:

As A/IS become more capable, as measured by the ability to perform with greater autonomy across a wider variety of domains, unanticipated or unintended behavior becomes increasingly dangerous.

Background

A/IS with an incorrectly or imprecisely specified objective function (or goals) could behave in undesirable ways (Amodei et al. 2016, Bostrom 2014, Yudkowsky 2008). 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) and Omohundro (2008) have argued that AGI systems are likely by default to adopt "convergent instrumental subgoals" such as resource-acquisition and self-preservation, unless the system is designed to explicitly disincentivize these strategies. These types of problems are likely to be more severe in systems that are more capable (as follows from their increased optimization power and broader action space range) unless action is taken to prevent them from arising.

In order to foster safety and controllability, A/IS 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 preemptively. When considering problems such as these, teams should cultivate a "safety mindset" (as described by Schneier [2008] in the context of computer security — to anticipate and preempt adversaries at every level of design and implementation), and suggest that many of these problems can likely be better understood by studying adversarial examples (as discussed by Christiano [2016]) and other A/IS robustness and safety research threads.

Teams working on such advanced levels of A/IS should pursue the following goals, all of which seem likely to help avert the above 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, Taylor et al. in Alignment for Advanced Machine Learning Systems, and Russell et al. in Research Priorities for Robust and Beneficial Artificial Intelligence. See also the work of Hadfield-Menell et al. (2016) and the references therein.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

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

2. Work to ensure that A/IS are transparent, i.e., that their internal reasoning processes can be understood by human operators. This likely involves both theoretical and practical research. In particular, teams should develop, share, and contribute to transparency and debugging tools that make the behavior of advanced A/IS easier to understand and work with; and teams should 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 infrastructure and environments for development, testing, and deployment of powerful A/IS. This work will provide some protection against risks including subversion by malicious external attackers, and unsafe behavior arising from exploratory learning algorithms. In particular, teams should develop, share, and contribute to AI safety test environments and tools and techniques for "boxing" A/IS (see Babcock et al. [2016] and Yampolskiy [2012] for preliminary work).

4. Work to ensure that A/IS "fail gracefully" (e.g., shutdown safely or go into some other known-safe mode) in the face of adversarial inputs, out-of-distribution errors (see Siddiqui et al. [2016] for an example), unexpected rapid capability gain, and other large context changes.

5. Ensure that A/IS are corrigible in the sense of Soares et al. (2015), i.e., that the systems are amenable to shutdown and modification by the operators, e.g., as with Hadfield-Menell (2017) and Russell et al. (2016), 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).

6. Explore methods for making A/IS capable of learning complex behaviors and goals from human feedback and examples, in spite of the fact that this feedback is expensive and sometimes inconsistent, e.g., as newer variants of inverse reinforcement learning attempt. See Evans et al. (2015) and Hadfield-Menell et al. (2016).

7. Build extensive knowledge layers and automated reasoning into systems to expand their contextual awareness and common sense so undesirable side effects can be determined and averted dynamically.

Candidate Recommendations

1. Teams working on developing AGI systems should be aware that many technical robustness and safety issues are even present in today's systems and that, given more research, some corrective techniques for those can likely scale with more complex problem manifestations.

2. Teams working on developing AGI systems should be prepared to put significantly more effort into AI safety research as capabilities grow.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

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

3. Teams working on developing AGI systems should cultivate a "safety mindset" like a "security mindset," vigilant of ways they can cause harm and invest in preventing those.

Issue:

Designing for safety may be much more difficult later in the design lifecycle rather than earlier.

Background

Different types of AGI systems are likely to vary widely in how difficult they are to align with the interests of their operators. As an example, consider the case of natural selection, which developed an intelligent "artifact" (brains) by a process analogous to a simple hill-climbing search algorithm. Brains are quite difficult to understand, and modifying a brain to be trustworthy when given large amounts of resources and unchecked power would be extremely difficult or impossible.

Similarly, systems developed using search/optimization, especially those using multiple layers of representations, might be difficult to modify/align. At the other end of the spectrum, we can imagine systems with more principled or explicit designs that are perfectly rational, understandable, and easy to modify/align. On this spectrum, a system like AlphaGo would be closer to the search/optimization/meta end of the spectrum, and Deep Blue closer to the other.

Realistic AGI systems are likely to fall somewhere in between, and will be built by a combination of human design and search/optimization (e.g., gradient descent, trial-and-error, etc.). Developing AGI systems without these concerns in mind could result in complicated systems that are difficult or impossible to align with the interests of its operators, leading to systems that are more vulnerable to the concerns raised above.

A relevant analogy for this issue is the development of the C programming language, which settled on the use of null-terminated strings instead of length-prefixed strings for reasons of memory efficiency and code elegance, thereby making the C language vulnerable to buffer overflow 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. Paying the upfront cost in this case would have prevented much larger costs that we are still paying today. (It does require skill though to envision the types of downstream costs that can result from upstream architectural changes.)

Given that some A/IS development methodologies will result in AGI systems that are much easier to align with intentions than other methodologies, and given that it may be quite difficult to switch development methodologies and architectures late in the development of a highly capable A/IS,

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

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

great care should be taken by teams developing systems intended to eventually reach AGI level to ensure that their development methodology, techniques, and architecture will result in a system that can be easily aligned. (See also the discussion of transparency tools above.)

As a heuristic, when teams develop potentially dangerous systems, those systems should 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] 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 do so due to restrictions on the system's action space, is not safe by design. Of course, all appropriate safety precautions should be used, 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.

Candidate Recommendation

When designing an advanced A/IS, researchers and developers should pay the upfront costs to ensure, to the extent possible, that their systems are "safe-by-design," and only use external restrictions on the system as fail-safes rather than as a first line of defense. This involves designing architectures using known-safe and more-safe technical paradigms as early in the lifecycle as possible.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

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 capable A/IS.

Background

Issues A/IS researchers and developers will encounter include challenges in determining whether a system will cause unintended and unanticipated harms — to themselves, the system's users, and the general public — as well as complex moral and ethical considerations, including even the moral weight of certain A/IS themselves or simulations they may produce (Sandberg 2014). Moreover, researchers and developers may be subject to cognitive biases that lead them to have an optimistic view of the benefits, dangers, and ethical concerns involved in their research.

Across a wide range of research areas in science, medicine, and social science, review boards have served as a valuable tool in enabling those with relevant expertise to scrutinize the ethical implications and potential risks of research activities. While A/IS researchers and developers themselves should be alert to such considerations, review boards can provide valuable additional oversight by fielding a diversity of disciplines and deliberating without direct investment in the advancement of research goals.

Organizations should set up review boards to support and oversee researchers working on projects that aim to create very capable A/IS. AI researchers and developers working on such projects should also advocate that these boards be set up (see Yampolskiy and Fox [2013] for a discussion of review boards for AI projects). There is already some precedent for this, such as Google DeepMind's ethics board (though not much is known publicly about how it functions).

Review boards should be composed of impartial experts with a diversity of relevant knowledge and experience. These boards should be continually engaged from the inception of the relevant project, 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 (international) standards or best practices developed by the industry/field as a whole, perhaps through groups like the Partnership on Artificial Intelligence, our IEEE Global Initiative, or per the Asilomar AI Principles.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

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

Review boards should be complemented by other measures to draw upon diverse expertise and societal views, such as advisory groups, relevant workshops and conferences, public engagement processes, and other forums for discussion and debate. The incorporation of a wide range of viewpoints, commensurate with the breadth and scale of potential impact, will support A/IS researchers and developers in making optimal design decisions without relying solely on the oversight of review boards.

Given the transformative impact AGI 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 team would face when deploying an AGI-level system, technologists should pursue AI designs that would bring about beneficial outcomes regardless of the moral fortitude of the research team. Teams should work to minimize the extent to which beneficial outcomes from the system hinge on the virtuousness of the operators.

Candidate Recommendation

1. Organizations working on sufficiently advanced A/IS should set up review boards to consider the implications of risk-bearing proposed experiments and development.

2. Technologists should work to minimize the extent to which beneficial outcomes from the system hinge on the virtuousness of the operators.

Issue:

Future A/IS may have the capacity to impact the world on a scale not seen since the Industrial Revolution.

Background

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

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

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

Furthermore, institutions should encourage A/IS researchers and developers, who are concerned that their lab or team is not following safety best practices, to raise this to the attention of the wider A/IS community without fear of retribution. Any group working to develop capable A/IS should understand that, if successful, their technology will be considered both extremely economically and politically significant. 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) and Krakovna (2016).

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). We encourage researchers and developers aspiring to develop these systems to take on this norm. It is imperative that the pursuit and realization of AGI systems be done in the service of the equitable, long-term flourishing of civilization.

In 2017, broad coalitions of AI researchers, ethicists, engineers, businesspeople, and social scientists came together to form and to endorse the Asilomar AI Principles (Future of Life Institute 2017), which includes the relevant principles "14) Shared Benefit: AI technologies should benefit and empower as many people as possible. ... 15) Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity. ... 23) Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization."

Candidate Recommendations

1. Adopt the stance that superintelligence should be developed only for the benefit of all of humanity.

2. De-stigmatize and remove other soft and hard barriers to AI researchers and developers working on safety, ethics, and beneficence, as well as being open regarding that work.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Autonomous and Intelligent systems (A/IS) are developing faster than the supporting standards and regulation required for transparency and societal protections can keep pace. The impact of these systems on society is direct and considerable.

A/IS require data to fuel learning, and inform automatic decision-making. Increasingly this data is personal data, or personally identifiable information, known as PII. PII is defined as any data that can be reasonably linked to an individual based on their unique physical, digital, or virtual identity. As a result, through every digital transaction (explicit or observed) humans are generating a unique digital shadow of their physical self.

Ethical considerations regarding data are often focused largely on issues of privacy — what rights should a person have to keep certain information to themselves or have input into how it is shared? However, individuals currently lack clarity around how to access, organize, and share their data to ensure unintended consequences are not the Laws are generally enforceable result. Without clarity, these issues will continue to reflect negatively on the proliferation of the A/IS industry.

The aim of this Committee is to set out the ethical considerations in the collection and use of personal data when designing, developing, and/or deploying A/IS. Furthermore, to entreat all global (A/IS) technologists (academics, engineers, programmers, manufacturers, and policy makers) to proactively prioritize and include individuals in the data processes that directly relate to their identity.

There is a fundamental need for people to have the right to define access and provide informed consent with respect to the use of their personal data (as they do in the physical world). Individuals require mechanisms to help curate their unique identity and personal data in conjunction with policies and practices that make them explicitly aware of consequences resulting from the bundling or resale of their personal information and life experiences.

Enabling individuals to curate their identities and manage the ethical implications of their data use will remain essential to human culture everywhere in the world. While some may

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

choose only minimum compliance to legislation like the European General Data Protection Regulation (GDPR), forward-thinking organizations will shift their data strategy (marketing, product, and sales) to enable methods of harnessing volunteered intentions from customers (or in governmental contexts, citizens), versus only invisibly tracking their attention or actions.

For individuals to be at the center of their data, policy makers and society at large will need to rethink the nature of standards and human rights as they have been applied to the physical world and to re-contextualize their application in the digital world. While standards exist, or are in production relating to augmented and virtual reality, human rights law, privacy and data, it is still largely not understood how human agency, emotion, and the legal issues regarding identity will be affected on a large scale by society once A/IS technologies become ubiquitous.

The goal of the analysis of these ethical issues and considerations by this Committee regarding data usage and identity is to foster a positive and inclusive vision for our shared future. To accomplish this goal, this document is focused on the following themes:

1. Digital Personas
2. Regional Jurisdiction
3. Agency and Control
4. Transparency and Access
5. Symmetry and Consent

We have also created an Appendix document listing key resources referenced in the following section.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Addressing these issues and establishing safeguards prioritizing the protection and assets of individuals regarding privacy and personal data in the realms of A/IS is of paramount importance today. To that end, since the creation of the first draft of *Ethically Aligned Design* this Committee recommended ideas for the following IEEE Standards Working Groups which have been and approved and are free for all to join (click on links for details):

- IEEE P7002™, Data Privacy Process
- IEEE P7004™, Standard on Child and Student Data Governance
- IEEE P7005™, Standard on Employer Data Governance
- IEEE P7006™, Standard for Personal Data Artificial Intelligence (AI) Agent

The goal of this Committee is that our recommendations, in conjunction with the development and release of these Standards once adopted, will expedite the prioritization and inclusion of all global individuals in the data processes that directly relate to their identity.

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

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Section 1 — Digital Personas

While many individuals may not currently have the ability to claim their identity (in the case of refugees, etc.), as a rule society understands how to apply the legal concepts of identity in real-life situations. In digital or virtual realms, however, our personas are fluid — individuals can be avatars in gaming situations or take on a different tone in various social networking settings. Behaviors regarding our personas considered normal in real-life are not directly applicable in the augmented, virtual and mixed reality worlds most individuals will soon be inhabiting on a regular basis in the near future. In regards to the algorithms powering AI, or the affective sensors becoming standard features in autonomous vehicles, or companion robots, etc., how A/IS affects our digital personas through use or misuse of our data is critical to understand, monitor, and control.

Issue:

Individuals do not understand that their digital personas and identity function differently than in real life. This is a concern when personal data is not accessible by an individual and the future iterations of their personas or identity cannot be controlled by them, but by the creators of the A/IS they use.

Background

A/IS created from personal experiences is different from AI created from farming or climate data. Society has had traditional safeguards on the use and application of personal information to encourage innovation and to protect minorities. Traditional systems for medicine and law limit secrecy and favor regulation of professionals at the edges over centralized hierarchical corporations. For example, almost 100% of intellectual property in the domains of medicine and law is open, peer-reviewable, and can be taught to anyone, anywhere.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

However, the emergence of the Internet of Things (IoT) and augmented reality/virtual reality (AR/VR) means personal information forms a foundation for every system being designed. This data acts as the digital representation and proxy for our identity. From birth, the different roles individuals take on in life provide specific contexts to the data they generate. Previously these contexts and roles enabled individuals to maintain some level of privacy due to the siloes of collection. Now, as the prospect of an omni-connected world approaches, those silos are being replaced by horizontal integrations that put the digital versions of personas and roles at risk. It is therefore important that citizens understand these roles and their related data to assess the downstream (further) consequences of its aggregation. Digital personas/roles include:

- Pre-birth to post-life digital records (health data)
- Birth and the right to claim citizenship (government data)
- Enrollment in school (education data)
- Travel and services (transport data)
- Cross-border access and visas (immigration data)
- Consumption of goods and services (consumer and loyalty data)
- Connected devices, IoT and wearables (telecommunications data)
- Social and news networks (media and content data)
- Professional training, internship, and work (tax and employment data)
- Societal participation (online forums, voting and party affiliation data)
- Contracts, assets, and accidents (insurance and legal data)
- Financial participation (banking and finance data)
- Death (digital inheritance data)

By the time individuals reach early adulthood, they are simultaneously acting across these roles, generating vast amounts of personal data that is highly contextual and easy to identify and link directly to an individual. If an individual's digital shadow is a proxy of their physical self, then technologists and policy makers must address the transparency, control, and asymmetry of how personal data is collected and used to enable A/IS. A/IS technologists need to recognize the coercive nature of many current identity schemes — such as hidden tracking by advertising brokers — and adopt privacy-preserving identity practices such as same-domain pseudonymous identifiers and self-sovereign identity.

Candidate Recommendation

The ethics of creating secret and proprietary A/IS from people's personally identifiable information (PII) need to be considered based on the potential impact to the human condition. To preserve human dignity, policies, protections, and practices must provide all individuals the same agency and control over their digital

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

personas and identity they exercise in their real-world iterations no matter what A/IS may be in place to monitor, assist, or interact with their data.

Further Resources

- Blockchain Identity (Rebooting Web-of-Trust).
- W3C Credentials Community Group.
- HIE of One.

Issue:

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

Background

Identity is emerging at the forefront of the risks and opportunities related to use of personal data for A/IS. Across the identity landscape there is increasing tension between the requirement for federated identities (all data linked to a natural and identified natural person) versus a range of identities (personas) that are context specific and determined by the use-case, for example opening a bank account, crossing a border, or ordering a product online. New movements, such as Self-Sovereign Identity — defined as the right of a person to determine his or her own identity — are emerging alongside legal identities (issued by governments, banks, and regulatory authorities) to help put individuals at the center of their data in the algorithmic age.

Personas (an identity that acts as a proxy) and pseudonymity are also critical requirements for privacy management since they help individuals select an identity that is appropriate for the context they are in or wish to join. In these settings, trust transactions can still be enabled without giving up the "root" identity of the user. For example, it is possible to validate a user is over 18 (for adult content) or eligible for a service (postcode confirmation). Attribute verification (comprising the use of empowered persona usage by an individual) will play a significant role in enabling individuals to select the identity that provides access without compromising agency. This type of access is especially important in dealing with the myriad algorithms interacting with data representing tiny representations of our identity where individuals typically are not aware of the context for how their data will be used.

Candidate Recommendation

Individuals should have access to trusted identity verification services to validate, prove, and support the context-specific use of their identity. Regulated industries and sectors such as banking, government, and telecommunications should provide data-verification services to citizens and consumers to provide greatest usage and control for individuals.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Further Resources

- The Inevitable Rise of Self-Sovereign Identity by The Sovrin Foundation.
- See Identity Examples in the Appendix Document for this section.
- IEEE P7006™, Standard for Personal Data Artificial Intelligence (AI) Agent Working Group. This Standards Working Group is free and open to anyone wishing to join and addresses issues relating to how an individual could have the ubiquitous and always-on services of a personalized AI agent to ensure their identity is protected and has symmetry with the A/IS their data comes into contact with at all times.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 2 — Regional Jurisdiction

Legislation regarding personal data varies widely around the world. Beyond issues of data operability issues when transferring between country jurisdictions, rights of individuals and their access and usage of data depends on the regions and laws where they live. Much of A/IS ethics involves the need to understand cultural aspects of the systems and services an organization wishes to create for specific users. This same attention must be given to how data related to A/IS are positioned from a regional perspective to best honor the use, or potential abuse of the global citizens' data. A/IS will also be subject to regional regulation, for example under the General Data Protection Regulation (GDPR), European citizens may have specific rights of redress where AI or AS has been used.

Issue:

Country-wide, regional, or local legislation may contradict an individual's values or access and control of their personal data.

Background

Ethical considerations regarding data are often focused largely on issues of privacy — what rights should a person have to keep certain information to themselves, or have input into how it is shared? While rhetoric in various circles stating, "privacy is dead" may be someone's personal opinion reflecting their values, privacy is nonetheless a fundamental human right recognized in the UN Declaration of Human Rights, the International Covenant on Civil and Political Rights, and in many other international and regional treaties.

However, this fundamental right is not universally recognized or supported. It is also culturally contextual and nuanced. It is therefore critical to understand the jurisdictional and specific legal requirements that govern the access and use of personal information when developing A/IS solutions. *These include, but are not limited to:*

- **Europe**; the introduction of the General Data Protection Regulation (GDPR), Personal Services Directive II (PSD2), and ePrivacy. These new regulations carry substantial fines for non-compliance. Depending on the nature and circumstances of the violation, these penalties may include:
  - A warning in writing in cases of first and non-intentional non-compliance
  - Regular periodic data protection audits

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

  - A fine up to 10,000,000 EUR or up to 2% of the annual worldwide turnover of the preceding financial year in case of an enterprise, whichever is greater (Article 83, Paragraph 4)
  - A fine up to 20,000,000 EUR or up to 4% of the annual worldwide turnover of the preceding financial year in case of an enterprise, whichever is greater (Article 83, Paragraph 5 and 6)

- **United States**: The United States lacks a single "baseline" privacy regime; instead, policies and procedures affecting the collection and use of PII varies based on type of information and which entity possesses the data. Laws, for example, afford certain procedural requirements around financial data, certain protected health information, and children's data. Laws are generally enforceable by state and federal regulators (including the Federal Trade Commission and state attorney general), though individuals may have private rights of action under state law or certain federal laws such as the Video Privacy Protection Act, which governs disclosures of identifiable video rental records, and the Fair Credit Reporting Act, which provides access and rights to consumer reports used for eligibility determinations. See also: Jurisdiction Examples in the Appendix Document for this section.
- **Australia**: In addition to strict privacy regulation, the Australian Productivity Commission issued reports in 2016 and 2017 acknowledging that personal information is a personal asset and therefore recognized the need for Australians to have control with respect to its collection and use. At the time of publication, The Australian Federal Government is in the process of using these reports to inform the drafting of new personal data regulation.
- **Japan**: The Act on the Protection of Personal Information was amended in 2016. The act precisely defines the definition of personal information; however, the concept of privacy is not explicitly stated. In this sense, the act is deemed as a practice-oriented law. The new concept of *anonymously processed information* is introduced which is produced to make it impossible to identify a specific individual. In addition, it can be transferred to, and used by, the third parties without the data subject's consent. The method of producing anonymously processed information will be determined on a sector-by-sector basis because each sector has distinct constraints and purposes of personal information.

Additionally, there is growing evidence that not providing clear consent (regarding personal data usage) decreases mental and emotional well-being. The rapid rise in ad blocking tools or lowering of consumer trust via reports of non-ethically driven online studies provides tangible evidence toward the failure of these clandestine efforts.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Candidate Recommendation

While specific uses of data must be taken in context of the regions where specific legislation applies, individuals should always be provided access to, and control of, their data to ensure their fundamental human rights are honored without fear of the risk of breaking applicable laws.

Further Resources

- Amended Act on the Protection of Personal Information in Japan.
- Outline of the Amended Personal Information Protection Act in Japan.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 3 — Agency and Control

Agency is the capacity of individuals to act independently and to exercise free choice, a quality fundamental to democratic ideals. Central to human agency is control. As society moves towards complete connectivity, humans will require tools and mechanisms to enable agency and control over how their personal data is collected and used. When people do not have agency over their identities political participation is impossible, and without political participation ethics will be decided by others. As the rise of algorithms accessing people's data relating to their identities continues, there is increased risk of loss of agency and well-being, adding the potential for depression and confusion along with the lack of clear ways to contribute ideas in an open and democratic fashion.

Issue:

To understand the role of agency and control within A/IS, it is critical to have a definition and scope of personally identifiable information (PII).

Background

Different laws and regulations around the globe define the scope of PII 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 be considered PII. This is further complicated by machine learning and autonomous systems that access and process data faster than ever before.

Multiple global bodies believe PII is a sovereign asset belonging to an identified individual. PII, or personal data, is defined as any data that can be reasonably linked to an individual based on their unique physical, digital, or virtual identity. PII protections are often related to the U.S. Fourth Amendment, as the right of the people to be secure in their persons, houses, papers, and effects.

As further clarification, the European Union definition of personal data set forth in the Data Protection Directive 95/46/ECl vi, defines personal data as "any information relating to an identified or identifiable natural person." Identifiable when? The question asked today will have a very different answer tomorrow given that all A/IS person-level or device-level data is identifiable if the tech advances and the data is still available. Agency requires that the control be exercised by the subject at the time the data is used, not at the time the data is collected.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Overall, personal data reflects self-determination and the inalienable right for an individual to be able to access and control the attributes of their physical, digital, and virtual identity.

Candidate Recommendation

Individuals should have access to means that allow them to exercise control over use of personal data at the time the data is used. If that agency and control is not available, person-level data needs to either be aggregated into larger cohorts and the person-level data deleted. PII should be defined as the sovereign asset of the individual to be legally protected and prioritized universally in global, local, and digital implementations regardless of whether deemed to be de-identified in the way it is stored.

Further Resources

- Determining What Is Personal Data, U.K. Information Commissioner's Office.
- Electronic Communications Privacy Act.
- Open PDS.
- IEEE Digital Inclusion through Trust and Agency Industry Connection Program.
- HIE of One — a patient-owned and controlled standards-based, open source EHR, so patients can collect, aggregate, and share their own data.

Issue:

What is the definition of control regarding personal data, and how can it be meaningfully expressed?

Background

Most individuals believe controlling their personal data only happens on the sites or social networks to which they belong, and have no idea of the consequences of how that data may be used by others tomorrow. Providing individuals with tools, like a personal data cloud, can empower users to understand how their data is an asset as well as how much data they produce. Tools like personal data vaults or clouds also let individuals organize their data around various uses (medical, social, banking). Control enables individuals to also assert a version of their own terms and conditions.

In the current context of A/IS technologies, and in the complex and multi-level or secondary uses of data, it is important to be clear about the boundaries of control for use of personal data that can affect an individual directly compared to collection of data for aggregated or systematic work (and exceptions for approved research). For example, an individual subway user's travel card, tracking their individual movements, should be protected from uses that identify or profile that individual to make inferences about his/her likes or location generally, but could be included in the overall travel systems management to

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

aggregate user data into patterns for scheduling and maintenance as long as the individual-level data is deleted.

The MyData movement combines related initiatives, such as Self Data, Vendor Relationship Management, Internet of Me, and Personal Information Management Systems (PIMS) under a common cause to empower individuals with their personal data. The Declaration of MyData Principles highlights human-centric control of personal data as one of core principles, emphasizing that people should be provided with the practical means to understand and effectively control who has access to data about them and how it is used and shared. In detail, the MyData Declaration states: "We want privacy, data security and data minimization to become standard practice in the design of applications. We want organizations to enable individuals to understand privacy policies and how to activate them. We want individuals to be empowered to give, deny or revoke their consent to share data based on a clear understanding of why, how and for how long their data will be used. Ultimately, we want the terms and conditions for using personal data to become negotiable in a fair way between individuals and organizations."

Candidate Recommendation

Personal data access and consent should be managed by the individual using systems that provide notification and an opportunity for consent at the time the data is used, versus outside actors being able to access personal data outside of an individual's awareness or control.

Further Resources

- Project VRM — vendor relationship management (VRM) tools and frameworks.
    - Kuan Hon, W. K., C. Millard, and I. Walden. "The Problem of 'Personal Data' in Cloud Computing — What Information Is Regulated? Cloud of Unknowing, Part 1." *Queen Mary School of Law Legal Studies Research Paper No. 75/2011; International Data Privacy Law* 1, no. 4 (2011): 211–228.
    - Boyd, E. B. "Personal.com Creates an Online Vault to Manage All Your Data." *Fast Company*, May 7, 2012._
- Meeco Life Management Platform. Personal cloud, attribute wallet and personal data management tools, consent engine and dual sided permission APIs.
    - MyData2017. Declaration of MyData Principles.
- Poikola, A. K. Kuikkaniemi, and H. Honko (Ministry of Transport and Communications). MyData — A Nordic Model for Human-Centered Personal Data Management and Processing. Finland: Prime Minister's Office, 2014.
- Hasselbalch, G., and P. Tranberg. "Personal Data Stores" (chapter 12), in Data Ethics: *The New Competitive Advantage*. Publishare, 2016.
- GDPR Article 20, Right to Data Portability, Article 29 Working Party, Brussels, 2016.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

- Thurston, B. "A Radical Proposal for Putting People in Charge of Their Data." *Fast Company*, May 11, 2015.
- de Montjoye, Y.-A., Wang, S. S., and Pentland, A. S. "openPDS: Protecting the Privacy of Metadata through SafeAnswers." *PLoS ONE* 9, no. 7 (2014): e98790.
- Definition of the *right* to be forgotten.
- IEEE Digital Inclusion through Trust and Agency. The Industry Connection Program develops comprehensive roadmaps, industry action reports, and educational platforms working to address issues around cyber-identity, digital personas, distributed ledger technology, and inclusion of underserved and vulnerable.
- See "The Attribute Economy 2.0," a multi-authored paper published by Meeco.
- The Path to Self-Sovereign Identity.
- uPort is an open source software project to establish a global, unified, sovereign identity system for people, businesses, organizations, devices, and bots. The Ethereum based self-sovereign identity system now in alpha testing.
- Sovrin—identity for all. The Sovrin Foundation describes self-sovereign identity (SSI) as "...an identity that is 100% owned and controlled by an individual or organization. No one else can read it, use it, turn it off, or take it away without its owner's explicit consent."
- Nichol, P. B. "A Look at India's Biometric ID System: Digital APIs for a Connected World." *CIO Perspectives*, February 23, 2017.
- See also Appendix 3: Digital Divide and Pay for Privacy.
- See also Appendix 4: Examples of Agency and Transparency.
- See also Appendix 5: Can Personal Data Remain Anonymous?

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 4 — Transparency and Access

Much of the contention associated with the concept of "privacy" actually relates to access. Challenges often arise around transparency and providing an explicit understanding of the consequences of agreeing to the use of people's personal data. This is 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. They obscure disclosure of the ways the data is used and make it hard to know what data was used. This can be especially evident via the invisible algorithms representing multiple services that access people's data long after they've provided original access to a service or their partners.

If individuals cannot access their personal data and account for how it is used, they cannot benefit from the insights that the data could provide. Barriers to access would also mean that individuals would not be able to correct erroneous information or provide the most relevant information regarding their lives to trusted actors. Transparency is also about notification. It is important that an individual is notified when their data is collected, and what usage is intended. In accordance with the GDPR, consent must be informed, explicit, and unambiguous.

Issue:

It is often difficult for users to determine what information a service provider or A/IS application collects about them at the time of such aggregation/collection (at the time of installation, during usage, even when not in use, after deletion). It is difficult for users to correct, amend, or manage this information.

Candidate Recommendation

Service providers should ensure that personal data management tools are easy to find and use within their service interface. *Specifically:*

- The data management tools should make it clear who has access to a user's data and for what purpose, and (where relevant) allow the user to manage access permissions.
- There should be legal, reputational, and financial consequences for failing to adhere to consent terms.
- It should be easy for users to remove their

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

data from the service. (Note: This is a GDPR requirement. It may not be mandated in the United States or for other services in countries outside of the EU, but represents a best-in-class practice to follow.) Organizations should create open APIs to their data services so that customers can access their data and governments should share the data they collect about their users directly with individuals and encourage them to ensure its accuracy for mutual value to combat the rising issue of dirty data.

Further Resources

- The User Managed Access Standard, proposed by The Kantara Initiative, provides a useful model to address these types of use cases.
- Surveys about how adults feel about health IT in 2005 and 2016 show that distrust of health technology has grown from 13% that withheld data from providers due to mistrust to 89%.

Issue:

How do we create privacy impact assessments related to A/IS?

Background

Because the ethical implications of intelligent systems are so difficult to discern, interested parties would benefit from analytical tools to implement standards and guidelines related to A/IS and privacy impacts. Like an environmental impact study or the GDPR privacy impact assessments, A/IS impact assessments would provide organizations with tools to certify their products and services are safe and consistent for the general public.

Candidate Recommendation

A system to assess privacy impacts related to A/IS needs to be developed, along with best practice recommendations, especially as automated decision systems spread into industries that are not traditionally data-rich.

Further Resources

In the GDPR in the EU, there is a requirement for a privacy impact assessment. The full report created by PIAF, The Privacy Impact Assessment Framework can be found here. In the report, of interest is Section 10.3, "Best Elements" whose specific recommendations provide insights into what could be emulated to create an AI impact assessment, including:

- PIA guidance documents should be aimed at not only government agencies but also companies or any organization initiating or intending to change a project, product, service, program, policy, or other initiative that could have impacts on privacy.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

- PIAs should be undertaken about any project, product, service, program, or other initiative, including legislation and policy, which are explicitly referenced in the Victoria Guide and the UK Information Commissioner's Office (ICO) Handbook.

Information privacy is only one type of privacy. A PIA should also address other types of privacy, e.g., of the person, of personal behavior, of personal communications, and of location.

- PIAF Consortium. "PIAF: A Privacy Impact Assessment Framework for Data Protection and Privacy Rights," 2011. Section 10.3.
- See the Personalized Privacy Assistant for a project applying these principles.
- While not explicitly focused on PIAs or AI, IEEE P7002™ Data Privacy Process is a Standards Working Group still open to join focused on these larger issues of data protection required by the enterprise for individuals' data usage.
- Usable Privacy Policy project for examples of how difficult privacy policies can be to maneuver.
- See also Appendix 4: Examples of Agency and Transparency.

Issue:

How can AI interact with government authorities to facilitate law enforcement and intelligence collection while respecting rule of law and transparency for users?

Background

Government mass surveillance has been a major issue since allegations of collaboration between technology firms and signals intelligence agencies such as the U.S. National Security Agency and the U.K. Government Communications Headquarters were revealed. Further attempts to acquire personal data by law enforcement agencies, such as the U.S. Federal Bureau of Investigation, have disturbed settled legal principles regarding search and seizure. A major source of the problem concerns the current framework of data collection and storage, which puts corporate organizations in custody of personal data and detached from the generators of that information. Further complicating this concern is the legitimate interest that security services have in trying to deter and defeat criminal and national security threats.

Candidate Recommendations

Personal privacy A/IS tools such as IEEE P7006™ have the potential to change the data paradigm and put the generators of personal information

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

at the center of collection. This would re-define the security services' investigative methods to pre-Internet approaches wherein individuals would be able to control their information while providing custody to corporate entities under defined and transparent policies.

Such a construct would mirror pre-Internet methods of information management in which individuals would deposit information in narrow circumstances such as banking, healthcare, or in transactions. This personal data AI agent would include root-level settings that would automatically provide data to authorities after they have satisfied sufficiently specific warrants, subpoenas, or other court-issued orders, unless authority has been vested in other agencies by local or national law. Further, since corporately held information would be used under the negotiated terms that the A/IS agent facilitates, authorities would not have access unless legal exceptions were satisfied. This would force authorities to avoid mass collection in favor of particularized efforts:

- The roots of the personal privacy A/IS should be devoid of backdoors that allow intrusion under methods outside of transparent legal authority. Otherwise, a personal A/IS could feed information to a government authority without proper privacy protection.
- Nuanced technical and legal techniques to extract warranted information while segregating and avoiding other information will be crucial to prevent overreach.
- Each request for data acquisition must come on a case-by-case basis versus an ongoing access form of access, unless the ongoing access has become law.
- Data-acquisition practices need to factor in the potential status of purely virtual representations of a citizen's identity, whether they do not have formal country of origin (physical) status, or their virtual identity represents a legal form of identity.
- Phasing in personal privacy AIs will mitigate risks while pre-empting reactive and disruptive legislation.
- Legal jurisdiction over personal privacy A/IS access will need to be clarified.

Further Resources

- UNECE. "Evaluating the Potential of Differential Privacy Mechanisms for Census Data." *Work Session on Statistical Data Confidentiality 2013*. Ottawa, October 28, 2013.
- CASD — Le Centre D'Accès Sécurisé Aux Données (The Secure Data Access Centre) is equipment that allows users, researchers, data scientists, and consultants to access and work with individual and highly detailed microdata, which are therefore subject to confidentiality measures, in the most secure conditions.
- Initiatives such as OPAL (for Open Algorithms), a collaborative project being developed by a group of partners committed

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

to leveraging the power of platforms, big data, and advanced analytics for the public good in a privacy-preserving, commercially sensible, stable, scalable, and sustainable manner.

- Ohm, P. "Sensitive Information." *Southern California Law Review* 88 (2015): 1125–1196.
- Y.-A. de Montjoye, L. Radaelli, V. K. Singh, A. S. Pentland. "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata." *Science* 347 (2015): 536–539.
- Sanchez, D., S. Martinez., and J. Domingo-Ferrer. "Comment on 'Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata'." Science 351, no. 6279 (2016): 1274–1274.
- Polonetsky, J., and O. Tene. "Shades of Gray: Seeing the Full Spectrum of Practical De-Identification." *Santa Clara Law Review* 56, no. 3 (2016): 593–629.
- Narayanan, A., and V. Shmatikov, "Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)." February 5, 2008.
- de Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen, and V. D. Blondel. "Unique in the Crowd: The Privacy Bounds of Human Mobility." *Scientific Reports* 3, no. 1376 (2013). doi: 10.1038/srep01376
- Coyne, A. "Government Pulls Dataset That Jeopardised 96,000 Employees." *iTnews*, October 6, 2016.
- Cowan, P. "Health Pulls Medicare Dataset After Breach of Doctor Details." *iTnews*, September 29, 2016.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 5 — Symmetry and Consent

Widespread data collection followed by the emergence of A/IS and other automated/autonomous data processing has placed tremendous strain on existing conceptions of "informed consent." This has created a vast asymmetry between the volume of organizations tracking individuals versus the tools allowing those individuals to fully understand and respond to all these tracking signals.

Legal frameworks such as the GDPR rely on the notion that data subjects must provide "freely given, specific, informed, and unambiguous" consent to certain data processing. Heavy reliance on a system of "notice and choice" has shifted the burden of data protection away from data processors and onto individual data subjects. A/IS can exacerbate this trend by complicating risk assessments of data sharing. When A/IS data transfer is done incorrectly it may alter or eliminate user interfaces, limiting choice and consent.

A/IS presents a new opportunity to offer individuals/end users a "real choice" with respect to how information concerning them is collected, used, and shared. Researchers are working to solve this issue in some contexts, but design standards and business incentives have yet to emerge.

Issue:

Could a person have a personalized privacy AI or algorithmic agent or guardian?

Background

For individuals to achieve and retain parity regarding their personal information in the algorithmic age, it will be necessary 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). The creation of personalized privacy A/IS would provide a massive opportunity for innovation in A/IS and corporate communities. There is natural concern that the rights of the individual are protected in the face of such opportunities.

The sophistication of data-sharing methodologies has evolved so these scenarios could 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" relationship — by allowing individuals to set their preferences for sharing and storing their data. An additional

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

benefit of finer-grained control of consent is that individuals are more likely to trust the organizations conducting research and provide more access to their data.

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 instance if a breach of conditions was detected.

Nonetheless, the dominant paradigm for personal data models needs to shift away from system and service-based models not under the control of the individual/human, and toward a model focused on the individual. 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 privacy A/IS or algorithmic guardian.

During the handshake/negotiation between the personal agent and the system or service, the personal agent would decide what data to make available and under what terms, and the system would decide whether to make the service available, and at what level. 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. A user should be able to override his/her personal agents should he/she decide that the service offered is worth the conditions imposed.

Vulnerable parts of the population will need protection in the process of granting access, especially given the asymmetry of power between an individual and entities.

Candidate Recommendations

Algorithmic guardian platforms should be developed for individuals to curate and share their personal data. Specifically:

1. 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.

2. For purposes of privacy, a person must be able to set up complex permissions that reflect a variety of wishes.

3. Default profiles, to protect naive or uninformed users, should provide little or no personal information without explicit action by the personal agent's owner.

4. The agent should help a person foresee and mitigate potential ethical implications of specific machine learning data exchanges.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

5. Control of the data from the agent should vest with the user, as otherwise users could lose access to his/her own ethical choices, and see those shared with third parties without permission.

6. A guardian should enable machine-to-machine processing of information to compare, recommend, and assess offers and services.

7. Institutional systems should ensure support and respect the ability for individuals to bring their own guardian to the relationship without any constraints that would make some guardians inherently incompatible or subject to censorship.

Further Resources

- The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Personal Data and Individual Access Control Section, in *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Artificial Intelligence and Autonomous Systems*, Version 1. IEEE, 2016.
- IEEE P7006™, Standard for Personal Data Artificial Intelligence (AI) Agent was launched in the summer of 2017 and is currently in development. Readers of this section are encouraged to join the Working Group if they are focused on these issues.
- 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.
- Privacy Assistant from MIT.

Issue:

Consent is vital to information exchange and innovation in the algorithmic age. How can we redefine consent regarding personal data so it respects individual autonomy and dignity?

Background

Researchers have long identified some key problems with notice and consent in the digital world. First, individuals cannot and will not read all of the privacy policies and data use statements to which they are exposed, and even if they could, these policies are not easy to understand.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Individual consent is rarely exercised as a meaningful choice due to poorly provisioned user-appropriate design.

A/IS place further strain on the notice and consent regime as further personalization of services and products should not be used as an excuse to minimize organizational transparency and choice for individuals to meet ethical and regulatory demand. If individuals opt not to provide personal information, they may find themselves losing access to services or receiving services based on stereotypes derived from the lower quality of data that they do provide.

When consent is not feasible or appropriate, organizations should engage in a robust audit process to account for processing of personal data against the interests of individuals. For instance, the GDPR permits processing on the grounds of an entity's legitimate interests, so long as those interests do not outweigh the fundamental rights and interests of data subjects. Organizations must develop internal procedures for conducting such an analysis, and external actors and regulators should provide further guidance and oversight where possible.

The needs of local communities, greater society, and public good should factor into this process. 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. This is particularly important where the primary reason for data collection may mask important secondary uses post-collection. In time, however, new mechanisms for facilitating dynamic consent rules and core structure as use-cases change. As data moves from the original collection context to a change of context, agile ethics rules should be deployed.

Candidate Recommendations

The asymmetric power of institutions (including public interest) over individuals should not force use of personal data when alternatives such as personal guardians, personal agents, law-enforcement-restricted registries, and other designs that are not dependent on loss of agency are available. When loss of agency is required by technical expedience, transparency needs to be stressed in order to mitigate these asymmetric power relationships.

Further Resources

- Office of the Privacy Commissioner of Canada. "Consultation on Consent Under the 'Personal Information Protection and Electronic Documents Act'." September 21, 2017. U.K. Information Commissioner's Office. "Consultation: GDPR Consent Guidance." March 2017.
- United Nations. "United Nations Declaration on the Rights of Indigenous Peoples." 107th plenary meeting, September 13, 2007.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Issue:

Data that is shared easily or haphazardly via A/IS can be used to make inferences that an individual may not wish to share.

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. As analysis becomes more obfuscated via A/IS, not even data controllers will necessarily know what or how conclusions are being drawn through the processing of personal data, or how those data are used in the whole process.

Opting out has some consequences. Users need to understand alternatives to consent to data collection before they give or withhold it, as meaningful consent. Without understanding the choices, consent cannot be valid. This places further strain on existing notions of informed consent. It raises the need for additional user controls and information access requirements. As computational power advances and algorithms compound existing data, information that was thought to be private or benign can be linked to individuals at a later time. Furthermore, this linked data may then be used to train algorithms, without transparency or consent, setting in motion unintended consequences. Auditing data use and collection for potential ethics risks will become increasingly more complex with A/IS in relation to these issues in the future.

Candidate Recommendation

The same A/IS that parses and analyzes data should also help individuals understand how personal information can be used. A/IS can prove granular-level consent in real time. Specific information must be provided at or near the point (or time) of initial data collection to provide individuals with the knowledge to gauge potential privacy risks in the long-term. Data controllers, platform operators, and system designers must monitor for consequences when the user has direct contact with an A/IS system. Positive, negative, and unpredictable impacts of accessing and collecting data should be made explicitly known to an individual to provide meaningful consent ahead of collection. Specifically:

- Terms should be presented in a way that allows the user to easily read, interpret, understand, and choose to engage with the system. To guard against these types of complexities, consent should be both conditional and dynamic. The downstream consequences (positive and negative) must be explicitly called out, such that the individual can make an informed choice, and/or assess the balance of value in context.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

- If a system impacts the ability of consumers to manage their own data via A/IS, accountability program management (PM) could be deployed to share consent solutions. A PM could span a diversity of tools and software applications to collect and transfer personal data. A PM can be assigned to evaluate consent metrics by ethics leadership to provide accountability reports. An actionable consent framework for personal data would not need to "reinvent the wheel." Existing privacy and personal data metrics and frameworks can be integrated into consent program management, as it becomes relevant. Likewise, resources, user controls, and policies should be put in place to afford individuals the opportunity to retract or erase their data if they feel it is being used in ways they do not understand or desire. Use limitations are also important and may be more feasible than collection limitations. At a minimum, organizations should commit to not use data to make sensitive inferences or to make important eligibility determinations absent consent. Because consent is so challenging in A/IS, it is vital that user participation, including data access, erasure, and portability, are also incorporated into ethical designs.
- Moving all computational values to the periphery (on the person) seems to be the only way to combat all the risks articulated. Systems should be designed to enable personalization and meta system learning concurrently without the permanent collection and storage of personal data for retargeting. This is a key architectural design challenge that A/IS designers must achieve if AI is going to be of service to society.

Further Resources

- Duhigg, C. "How Companies Learn Your Secrets." *The New York Times Magazine*, February 19, 2012.
- Meyer, R. "When You Fall in Love, This Is What Facebook Sees." The Atlantic, February 15, 2014.
- Cormode, G. "The Confounding Problem of Private Data Release." 18th International Conference on Database Theory (2015): 1–12.
- 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.
- OECD Standard of Data Minimization — Minimum data required for maximum service.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Issue:

Many A/IS will collect data from individuals they do not have a direct relationship with, or the systems are not interacting directly with the individuals. How can meaningful consent be provided in these situations?

Background

Individuals can be better informed of uses, processing, and risks of data collection when they interact with a system. IoT presents evolving challenges to notice and consent. Data subjects may not have an appropriate interface to investigate data controller uses and processes. They may not be able to object to collection of identifiable information, known or unknown to them by wireless devices, driven by A/IS.

When individuals do not have a relationship with the data collecting system, they will have no way of participating in their data under the notice and consent regime. This challenge is frequently referenced as the "Internet of Other People's Things." A/IS embodied in IoT devices and value-chains will need better interfaces and functionality to help subjects understand and participate in the collection and use of their data.

Candidate Recommendations

Where the subject does not have a direct relationship with the system, consent should be dynamic and must not rely entirely on initial terms of service or other instruction provided by the data collector to someone other than the subject. A/IS should be designed to interpret the data preferences, verbal or otherwise, of all users signaling limitations on collection and use, discussed further below.

Further Resources

- Kaminski, M. "Robots in the Home: What Will We Have Agreed To?" Idaho Law Review 51, no. 661 (2015): 551–677.
- Jones, M. L. "Privacy Without Screens and the Internet of Other People's Things," Idaho Law Review 51, no. 639 (2015): 639–660.
- Cranor, L. F. "Personal Privacy Assistants in the Age of the Internet of Things," presented at the World Economic Forum Annual Meeting, 2016.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Issue:

How do we make better user experience and consent education available to consumers as standard to express meaningful consent?

Background

Individuals are often not given agency or personal tools to express, invoke, or revoke consent to the terms of service or privacy and/or data use policies in their contracts. In many cases, individual data subjects were not notified at all of the transfer of their data in the course of business or government exchanges.

Industry data uses have led to individual exposure to intangible and tangible privacy harms, for example, mistaken identity. Inability to manage or control information has also led to barriers to employment, healthcare, and housing. This dynamic has resulted in some consumer resignation over the loss of control over personal information, despite a stated desire for additional control.

Candidate Recommendations

Tools, settings, or consumer education are increasingly available and should be utilized to develop, apply, and enforce consumer consent. *Specifically:*

- **Design the terms of service (ToS) as negotiable to consumers** — Combine user interface design to control the rate and method of data exchange, and provide a corporate terms ombudsman staffed as human agency to consumers facing a terms of service contract. Software developers would produce contract management platforms appropriate for consumer negotiation. This would support features to negotiate terms of consent contracts fairly for meaningful consumer consent. An example metric would be a consumer agreement held to 85% of a terms of service agreement content, as grounds to move forward with the contract. Companies conclude what the "deal breakers" or non-negotiables are ahead of time.
- **Provide "privacy offsets" as a business alternative to the personal data exchange** — Provide a pay alternative to the freemium data exchange model, to limit or cap third party vendor access to personal data or limit transactional data to internal business use only. Business developers would have to cost count individual data based on a general market profile, or offer a flat rate for advertising-free service. If they know immediately how much money they will lose if a new user would not consent to an external data exchange, they have grounds to pass the cost to new consumers as a privacy offset product.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

- **Apply "consent" to further certify artificial intelligence legal and as ethics doctrine** — Legal consent principles could be applied to a larger self-regulatory or co-regulatory artificial intelligence ethics certification framework for businesses and governments. This would be similar to medical certifications in ethics as a professional requirement, supportive of the Hippocratic Oath. Artificial intelligence ethics certification for responsible institutions (medical, government, education, corporations) should include education in applied legal consent principles, situation training regarding forms of consent, ethics certification testing, and perhaps a notarized public declaration to uphold ethical principles of consent. As an ethics board is formed it might: evaluate complaints, resolve ethical conflicts related to artificial intelligence and consent issues, improve upon current ethics procedures for consent, request independent investigations, review licensure or certification determinations, recommend professional penalties or discipline to organizations, and/or file legal claims based on findings.

- **Aggregate and provide visualization options for terms of service and privacy statements** — One way to provide better education and improved user experience, with respect to legal terms of use, is to offer visual analytics tools as a consumer control point of reference. Potential examples of this sort of effort include the Terms of Service Didn't Read Project and the Clarip. Both tools simplify the content of these policies and may provide users with clarity into how services are collecting, making use of, and potentially sharing personal and other information.

Further Resources

- Cavoukian, A. "Privacy by Design: The 7 Foundational Principles. Implementation and Mapping of Fair Information Practices." Internet Architecture Board, 2010.
- "From Consent to Data Control by Design." *Data Ethics*, March 20, 2017.
- Hintze, M. *Privacy Statements: Purposes, Requirements, and Best Practices.* Cambridge, U.K.: Cambridge University Press, 2017.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Issue:

In most corporate settings, employees do not have clear consent on how their personal information (including health and other data) is used by employers. Given the power differential between employees and employers, this is an area in need of clear best practices.

Background

In the beginning stages of onboarding, many employees sign hiring agreements that license or assign the usage of their data in very non-specific ways. This practice needs to be updated, so that it is clear to the employee what data is collected, and for what purpose. The employee must also have the ability/possibility to request privacy for certain data as well as have the right to remove the data if/when leaving the employment.

Candidate Recommendation

In the same way that companies are doing privacy impact assessments for how individual data is used, companies need to create *employee data impact assessments* to deal with the specific nuances of corporate specific situations. It should be clear that no data is collected without the consent of the employee.

Furthermore, it is critical that the data:

- Is gathered only for specific, explicitly stated, and legitimate purposes
- Is correct and up to date
- Is only processed if it is lawful
- Is processed in a proper manner, and in accordance with good practice
- Is not processed for any purpose that is incompatible with that for which the data was gathered
- Is rectified, blocked, or erased if it is incorrect or incomplete having regard for the purpose of the processing
- Is not kept for a longer period than is necessary

Further Resources

- The Swedish Personal Data Protection Act is taking a generic approach to data protection and data privacy, but it is well applicable for the specific case of employee data.
- IEEE P7005™, Standard for Transparent Employer Data Governance. *This Working Group is open and free for anyone to join.*

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Personal Data and Individual Access Control

Issue:

People may be losing their ability to understand what kinds of processing is done by A/IS on their private data, and thus may be becoming unable to meaningfully consent to online terms. The elderly and mentally impaired adults are vulnerable in terms of consent, presenting consequence to data privacy.

Background

The poor computer literacy of the elderly has been well known from the beginning of the information and Internet age. Among various problems related to this situation, is the financial damage caused by the misuse of their private information, possibly by malicious third parties. This situation is extremely severe for elderly people suffering from dementia.

Candidate Recommendations

- Researchers or developers of A/IS have to take into account the issue of vulnerable people, and try to work out an A/IS that alleviates their helpless situation to prevent possible damage caused by misuse of their personal data.
- Build an AI advisory commission, composed of elder advocacy and mental health self-advocacy groups, to help developers produce a level of tools and comprehension metrics to manifest meaningful and pragmatic consent applications.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

Autonomous systems designed to cause physical harm have additional ethical dimensions as compared to both traditional weapons and autonomous systems not designed to cause harm. Multi-year discussions on international legal agreements around autonomous systems in the context of armed conflict are occurring at the United Nations (UN), but professional ethics about such systems can and should have ethical standards covering a broad array of issues arising from the automated targeting and firing of weapons.

Broadly, we recommend that technical organizations promote a number of measures to help ensure that there is meaningful human control of weapons systems:

- That automated weapons have audit trails to help guarantee accountability and control.
- That adaptive and learning systems can explain their reasoning and decisions to human operators in transparent and understandable ways.
- That there be responsible human operators of autonomous systems who are clearly identifiable.
- That the behavior of autonomous functions should be predictable to their operators.
- That those creating these technologies understand the implications of their work.
- That professional ethical codes are developed to appropriately address the development of autonomous systems and autonomous systems intended to cause harm.

Specifically, we would like to ensure that stakeholders are working with sensible and comprehensive shared definitions, particularly for key concepts relevant to autonomous weapons systems (AWS). Designers should always ensure their designs meet the standards of international humanitarian law, international human rights law, and any treaties or domestic law of their particular countries, as well as any applicable engineering standards,

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

military requirements, and governmental regulations. We recommend designers not only take stands to ensure meaningful human control, but be proactive about providing quality situational awareness to operators and commanders using those systems. Professional ethical codes should be informed by not only the law, but an understanding of both local- and global-level ramifications of the products and solutions developed. This should include thinking through the intended use or likely abuse that can be expected by users of AWS.

While the primary focus of this document is with kinetic AWS that cause physical harm, it is recognized that many of these concerns and principles may also apply to cyber-weapons. This is, of course, also pertinent to cyber-weapons that have kinetic effects, such as those that destroy civilian infrastructures or turn civilian objects, vehicles, or infrastructure into kinetic weapons.

Additionally, society must be aware of the variety of political and security threats posed by AWS. Miniaturized AWS will pose additional threats because they are small, insidious, or obfuscated, and may therefore be non-attributable to the deploying entity. Depending upon payload or weapons (such as chemical, biological, or nuclear weapons), these may autonomously deploy weapons of mass destruction (WMD), or themselves constitute a new form of WMD. Additional ethical recommendations are needed to prevent the development of systems having these dangerous properties.

- Issues 1–3 raise general high-level questions regarding the definition of AWS and their relation to existing law and ethics.
- Issues 4–10 raise socio-political concerns over the likely uses and effects of AWS development and use.
- Issue 11 raises engineering concerns over the specific challenges posed by autonomous systems capable of targeting and deploying weapons.

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

Issue 1:

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 AWS definitions is not just an academic concern. The lack of clear definitions regarding what constitutes AWS is often cited as a reason for not proceeding toward any kind of international governance over autonomous weapons. 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 term *autonomy* is important for understanding debates about AWS; yet there may be disputes — about what the term means and whether what the definition identifies is technically possible today. This prevents progress in developing appropriate policies to regulate AWS design, manufacture, and deployment. Consistent and standardized definitions are needed to enable effective discussions of AWS, but they should be general enough to enable flexibility to ensure that those definitions do not become quickly technologically outdated.

Moreover, the phrases "human in the loop" and "human on the loop" also lack clarity and only contribute further confusion. Depending upon what one means, "in the loop" or "on the loop" means different things to different people. It could be used to describe the command chain that authorizes weapon release, where the commands flow down to a human and a weapon system to take specific actions. Yet, there are micro-level decisions where a human operator may have an opportunity to question the command. What often matters is the time delay between the fielding of an autonomous system, the decision to engage a weapon against a target, and the impact time.

Contrarily, "in the loop" obscures another temporal question: that whether in these scenarios clearance to fire at a target entails an authorization to prosecute that target indefinitely, or whether there are necessarily predetermined limits on the amount of time or ordinance each clearance provides. Central to this issue is how long a target that has been designated and verified by an authorized human in a given situational context remains a legitimate target.

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 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.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

Candidate Recommendations

The term *autonomy* in the context of AWS should be understood and used in the restricted sense of the 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 control of specific functions delegated to a given system, as well as the ways in which control over those functions are shared between human operators and AWS.

We support the working definition of AWS offered by the International Committee of the Red Cross (ICRC) and propose that it be adopted as the working definition of AWS for the further development and discussion of ethical standards and guidelines for engineers. The ICRC defines an AWS as: "any weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention."

Further Resources

- Dworkin, G. *The Theory and Practice of Autonomy.* Cambridge, U.K.: Cambridge University Press, 1988.
- Frankfurt, H. G. "Freedom of the Will and the Concept of a Person," in *The Importance of What We Care About*, Cambridge, U.K.: 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, R. *Autonomy: Beyond Negative and Positive Liberty.* New York: St. Martin's Press, 1986.
- Society of Automotive Engineers. J3016, Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems. SAE International, 2014.
- Roff, H. M. "An Ontology of Autonomy: Autonomy in Weapons Systems," in *The Ethics of Autonomous Weapons*, edited by C. Finkelstein, D. MacIntosh, and J. D. Ohlin. Cambridge, U.K.: Oxford University Press, forthcoming.
- Sharkey, N. "Towards a Principle for the Human Supervisory Control of Robot Weapons." *Politica and Società* 2 (2014): 305–324.
- U.K. Ministry of Defence. UK Joint Doctrine Note (JDN) 3/10, "Unmanned Aircraft Systems: Terminology, Definitions and Classification." May 2010.
- U.K. Ministry of Defence. UK Joint Doctrine Note (JDN) 2/11, "The UK Approach to Unmanned Aircraft Systems." March 2011.
- United Nations Institute for Disarmament Research (UNIDIR). "Framing Discussions on the Weaponization of Increasingly Autonomous Technologies." 2014
- International Committee of the Red Cross (ICRC). "Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons." September 1, 2016.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

Issue 2:

The addition of automated targeting and firing functions to an existing weapon system, or the integration of components with such functionality, or system upgrades that impact targeting and automated weapon release should be considered for review under Article 36 of Additional Protocol I of the Geneva Conventions.

Background

According to Article 36 of Additional Protocol I to the Geneva Conventions (1977), "In the study, development, acquisition or adoption of a new weapon, means or methods of warfare," weapon systems must be internally reviewed for compliance with international humanitarian law (IHL). Alterations to the critical functions or targeting and weapons release of an already-reviewed weapons systems should be considered for review, and any system automating those functions should be reviewed to ensure meaningful human control.

International human rights law (IHRL) also guarantees, by way of international and bilateral treaties, rights to life, human dignity, fair trial, and further positive and negative human rights. Society and engineers must consider the ways in which these rights may be threatened by the deployment and/or use of AWS, during armed conflict, policing, or other security operations.

There are situational and operational limitations of all engineered systems, and complete knowledge is not something that can be expected or required. However, there must be a multi-level effort to:

- Evaluate the conformity of a system to the law
- Evaluate its reliability and applicability for a given mission
- Evaluate its ability to conform to rules of engagement

Further, key decision makers need to understand the engineering constraints and limitations of weapons systems with high degrees of autonomy.

Candidate Recommendations

- All engineering work should conform to the requirements of international law, including both IHL and IHRL, as well as national and local laws. While this is not the primary responsibility of an individual engineer, there ought to be opportunities for engineers to learn about their obligations, their responsibilities with respect to AWS, as well as keeping their employing agencies accountable.
- Meaningful human control over the critical functions in weapons systems can help ensure that weapons can be used in conformity with the law in each instance. It is

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

also necessary for all stakeholders to consider design and implement accountability measures to help ensure all weapons are used in conformity with the law.

- Engineering constraints should be clearly identified, defined, and communicated to Article 36 weapons reviewers, to operators in their training for a system, and to military commanders and their legal counsel charged with specifying the rules of engagement.
- All those with responsibilities for weapon systems should ensure that Article 36 reviews will be held and provide all evidence needed at them. This should include any data which will lead to restrictions on their use, which will also be needed for Article 36 reviews and for military staff to set rules of engagement for the weapon system's use.
- There should be greater engineering input into the weapons reviews, and greater communication between engineers and lawyers in the weapons review process to ensure meaningful human control over weapons.

Further Resources

- International Committee of the Red Cross (ICRC). "Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons." September 1, 2016.

Issue 3:

Engineering work should conform to individual and professional organization codes of ethics and conduct. However, existing codes of ethics may fail to properly address ethical responsibility for autonomous systems, or clarify ethical obligations of engineers with respect to AWS. Professional organizations should undertake reviews and possible revisions or extensions of their codes of ethics with respect to AWS.

Background

- The ethical requirements for engineering have an independent basis from the law, although they are hopefully aligned with written laws and written codes of professional ethics. Where agreed upon, ethical principles are not reflected in written laws and ethical codes, individuals and organizations should strive to correct those gaps.
- Ethical requirements upon engineers designing autonomous weapon systems may go beyond the requirements of meeting local, national, and international laws.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

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.

Organizations such as the IEEE, the Association for Computing Machinery (ACM), the Association for the Advancement of Artificial Intelligence (AAAI), the UK Royal Academy of Engineering, the Engineering Council, Engineers Canada, and the Japanese Society for Artificial Intelligence (JSAI) have developed codes of ethics. Some of these groups are currently reviewing those codes in light of current and future developments in autonomous systems and AI.

While national laws may differ on what constitutes responsibility or liability for the design of a weapon 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 an individual is morally responsible requires understanding the organizations in which they work and to establish relevant facts in relation to the individual's acts and intentions.

Candidate Recommendations

Codes of conduct should 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 toward 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. Codes of conduct should also 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 may contravene (IHL) or (IHRL).

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 her contribution will enable or give rise to AWS and their use cases. Members must be aware

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

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 weapon system that is under meaningful human control — this may act as a justification for their research.

Further Resources

- Kvalnes, Ø. "Loophole Ethics," in *Moral Reasoning at Work: Rethinking Ethics in Organizations*, 55–61. Palgrave Macmillan U.K., 2015.
- Noorman, M. "Computing and Moral Responsibility," *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta , Summer 2014 Edition.
- Hennessey, M. "Clearpath Robotics Takes Stance Against 'Killer Robots'." Clearpath Robotics, 2014.
- "Autonomous Weapons: An Open Letter from AI & Robotics Researchers." Future of Life Institute, 2015.
- Noorman, M. "Computing and Moral Responsibility," in *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), edited by Edward N. Zalta.
- "Engineers Canada Code of Ethics," 2017.
- The Japanese Society for Artificial Intelligence Ethical Guidelines, 2017
- Engineering Council and Royal Academy of Engineering, Statement of Ethical Principles for the Engineering Profession.

Issue 4:

The development of AWS by states is likely to cause geopolitical instability and could lead to arms races.

Background

The widespread adoption of AWS by nation states could present a unique risk to the stability of international security. Because of the advantages of either countering an adversary through concomitant adoption of arms or being the first or prime mover is an offset advantage, the pursuit of AWS is likely to spur an international arms race. Evidence of states seeking greater adoption of artificial intelligence and quantum computing for security purposes already exists. The deployment of machine learning and other artificial intelligence applications on weapons systems is not only occurring, but will continue to advance. Thus it is important to look to previous scholarship on arms race dynamics to be informed about the first- and second-order effects of these races, such as the escalatory effects, arms development, decreasing international stability, and arms proliferation.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

Candidate Recommendations

Autonomous weapons designers should support the considerations of the United Nations to adopt a protocol to ensure meaningful human control over AWS under the Convention on Certain Conventional Weapons (CCW) treaty, or other similar effort by other international bodies seeking a binding international treaty.

It is unethical to design, develop, or engineer AWS without ensuring that they remain reliably subject to meaningful human control. Systems created to act outside of the boundaries of "appropriate human judgment," "effective human control," or "meaningful human control," violate fundamental human rights and undermine legal accountability for weapons use. Various scenarios for maintaining meaningful human control over weapons with autonomous functions should be further investigated for best practices by a joint workshop of stakeholders and concerned parties (including, but not limited to, engineers, international humanitarian organizations, and militaries), and that those best practices be promoted by professional organizations as well as by international law.

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.
- Roff, H. M., and R. Moyes. "Meaningful Human Control, Artifical Intelligence and Autonomous Weapons." Briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, April 2016.
- United Nations Institute for Disarmament Research (UNIDIR). "The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward." 2014.

Issue 5:

The automated reactions of an AWS could result in the initiation or escalation of conflicts outside of decisions by political and military leadership. AWS that engage with other AWS could escalate a conflict rapidly, before humans are able to intervene.

Background

One of the main advantages cited regarding autonomous weapons is that they can make decisions faster than humans, enabling rapid defensive and offensive actions. When opposing autonomous weapons interact with each other, conflict might escalate without explicit human military or political decisions, and escalate more quickly than humans on either side will be able to understand or act.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

Candidate Recommendations

- Consider ways of limiting potential harm from automated weapons. 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. All such solution options ought to precede the design, development, deployment, and use of weapons systems with automated targeting and firing functions.
- Perform further research on how to temper such dynamics when designing these systems.

Further Resources

- Scharre, P. "Autonomous Weapons and Operational Risk." Washington, DC: Center for New American Security, February, 2016.

Issue 6:

There are multiple ways in which accountability for the actions of AWS 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 maintain meaningful human control?), 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?).

Ideally all procurers, suppliers, and users of weapons systems components have accountability for their part of every weapons system, potential incorporation in future systems, and expected and potential users.

Candidate Recommendations

- Designers should follow best practices in terms of design process, which entails clearly defined responsibilities for organizations, companies, and individuals within the process.
- Systems and components should be designed to deter the easy modification of the overall weapon after the fact to operate in fully autonomous mode.
- Further exploration of black box recording of data logs, as well as cryptographic, block-chain, and other technical methods for tracing access and authorization of weapons targeting and release is needed.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

- System engineers must work to the same high standards and regulations of security for AWS design from a cybersecurity perspective than they would for any other work. Weapons systems ought to be designed with cybersecurity in mind such that preventing tampering, or at least undetected tampering, is a highly weighted design constraint.
- Procurement authority: only contract with contractors who have proper legal and security processes; carry out Article 36 reviews at all major steps in the procurement process; maintain database of design, tests, and review evidence.
- Contractors: ensure design meets relevant engineering and defense standards for military products; deliver evidence for Article 36 reviews using, but not restricted to, design reviews and simulation models; provide evidence requested by user for setting ROE; ensure design has clear criteria for decisions made by their product.
- Acceptance body: have validation and test plans for behavior of actual system produced; test weapons systems in a number of representative scenarios; have plans to ensure upgrades are reviewed against IHL criteria such as Article 36.
- User/military commanders: only operate weapons systems with meaningful human control and in accordance with delegated authority.
- Weapons systems must have default modes of operation agreed with campaign planners before operation commences.
- Ensure as many aspects of weapons systems as possible are designed with fail-safe behaviors.
- Ensure clear embedded lines of accountability in the design, deployment, and operation of weapons.
- Trusted user authentication logs and audit trail logs are necessary, in conjunction with meaningful human control. Thorough human-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 operational hardware.

There will need to be checks that all these bodies and organizations have discharged their responsibilities according to IHL and their domestic laws. Even if this is the case, weapons system operations may be compromised by, for example, equipment failure, actions by

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

opponents such as cyber-attacks, or deception so that the automated functions act according to design but against an incorrect target.

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 SEA-RAM, 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 U.S. Department of Defense (DoD) 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. However, most of these systems are only utilized in a defensive posture for close-in weapons systems support against incoming lethal threats.

Further Resources

- Gubrud, M., and J. Altmann. "Compliance Measures for an Autonomous Weapons Convention." International Committee for Robot Arms Control, 2013.
- U.K. Ministry of Defence. "The UK Approach to Unmanned Aircraft Systems (UAS)," Joint Doctrine Note 2/11, March 2011.
- Sharkey, N. "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.
- Roff, H. M., and R. Moyes. "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons Systems." Briefing Paper for the Delegates at the Convention on Certain Conventional Weapons Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, April 2016.
- Roff, H. M. "Meaningful Human Control or Appropriate Human Judgment." Briefing Paper for the Delegates at the 5th Review Conference at the Convention on Certain Conventional Weapons, Geneva, December 2016.
- Scherer, M. "Who's to Blame (Part 4): Who's to Blame if an Autonomous Weapon Breaks the Law?" *Law and AI*, February 24, 2016.
- Rebecca C, "War Torts: Accountability for Autonomous Weapons." *University of Pennsylvania Law Review* 164, no. 6 (2016): 1347–1402.
- Gillespie, T., and R. West. "Requirements for Autonomous Unmanned Air Systems Set by Legal Issues." *International C2 Journal* 4, no. 2 (2010): 1–32.
- Defense Science Board. "Summer Study on Autonomy." Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016.
- Rickli, J.-M. "Artificial Intelligence and the Future of Warfare" (Box 3.2.1). *2017 Global Risk Report,* Geneva: World Economic Forum, 2017.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

Issue 7:

AWS offer the potential for severe human rights abuses. Exclusion of human oversight from the battlespace can too easily lead to inadvertent violation of human rights. AWS could be used for deliberate violations of human rights.

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.

The potential lack of clear lines of accountability for the consequences of AWS might encourage malicious use of AWS by those seeking to avoid responsibility for malicious or illegal acts.

Candidate Recommendations

Acknowledge that the design, development, or engineering of AWS for anti-personnel or anti-civilian 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 to be 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 intended to increase, rather than decrease, situational awareness.

Under no circumstances is it morally permissible to use AWS without meaningful human control, and this should be prohibited. Ultimately, weapons systems must be under meaningful human control. 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

- Heller, K. J. "Why Preventive Self-Defense Violates the UN Charter." *Opinio Juris*, March 7, 2012.
- Scherer, M. "Who's to Blame (Part 5): A Deeper Look at Predicting the Actions of Autonomous Weapons." *Law and AI*, February 29, 2016.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

- Roff, H. M. "Killer Robots on the Battlefield: The Danger of Using a War of Attrition Strategy with Autonomous Weapons." *Slate,* 2016.
- Roff, H. "Autonomous Weapons and Incentives for Oppression." *Duck of Minerva,* March 13, 2016.

Issue 8:

AWS could be used for covert, obfuscated, and non-attributable attacks.

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.

Features of AWS that may contribute to their making covert and non-attributable attacks easier include: small size; the ability to swarm; and ability to act at great distance and time from the deployment of a weapon from responsible operators; layers of weapons systems within other systems.

States have a legal obligations to make attacks practically attributable. There are additional legal obligations not to booby trap autonomous systems. Self-destructive functions, such as those aimed at preventing access to sensitive technologies or data, should be designed to not cause incidental or intentional harm.

There are significant concerns about the use of AWS by non-state actors, or individuals, and the potential for use in terror attacks against civilians, and non-attributable attacks against states. Designers should be concerned about the potential of systems to be used by malicious actors.

Candidate Recommendation

Because AWS are delegated authority to use force in a particular situation, they are required to be attributable to the entity and human that deployed them. Designers should ensure that there is a clear and auditable authorization of actions taken by the AWS when in operation.

Further Resources

- Bahr, E. "Attribution of Biological Weapons Use," in *Encyclopedia of Bioterrorism Defense.* Hoboken, NJ: John Wiley & Sons, 2005.
- Mistral Solutions. "Close-In Covert Autonomous Disposable Aircraft (CICADA) for Homeland Security," 2014.
- Piore, A. "Rise of the Insect Drones." *Popular Science*. January 29, 2014.
- Gillespie, T., and R. West. "Requirements for Autonomous Unmanned Air Systems Set by Legal Issues." *International C2 Journal* 4, no. 2 (2010): 1–32.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

Issue 9:

The development 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 appearance.

Candidate Recommendations

- One must design weapons with high degrees of automation 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 10:

AWS could be deployed by domestic police forces and threaten lives and safety. AWS could also be deployed for private security. Such AWS may have very different design and safety requirements than military AWS.

Background

Outside of military uses of AWS, other likely applications include use by domestic police forces, as well as coast guards, border patrols, and other domestic security applications. Police in Dallas, Texas used a bomb disposal robot to deliver a bomb to kill a suspect in the summer of 2016. While that was a remotely operated weapon delivered by a remote operated platform, the path to more autonomous forms of police robots using weapons seems highly likely.

Beyond use by governments, AWS could potentially also be deployed for other private

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

security applications, such as guarding property, patrolling areas, and personal protection.

Tyrants and despots might utilize AWS to gain or retain control over a population which would not otherwise support them. AWS might be turned against peaceful demonstrators when human law enforcement might not do the same.

Candidate Recommendations

- Police and private security systems should not be permitted to deploy weapons without meaningful human control.
- Police and security systems should deploy non-lethal means to disrupt and avert security threats and threats to the physical safety of humans.

Further Resources

- Asaro, P. "Will #BlackLivesMatter to RoboCop?" WeRobot 2016, University of Miami School of Law, Miami, FL, April 1–2, 2016.
- Asaro, P. "'Hands Up, Don't Shoot!' HRI and the Automation of Police Use of Force," Special Issue on Robotics Law and Policy, *Journal of Human-Robot Interaction* 5, no. 3 (2016): 55–69.

Issue 11:

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

Background

Autonomous systems that react and adapt to environmental and sensor inputs results in systems that may be predictable in their general behavior, but may manifest individual or specific actions that cannot be predicted in advance.

As autonomous systems become more complex in their processing of data, the ability of designers to anticipate and predict their behavior becomes increasingly difficult.

As adaptive systems modify their functional operations through learning algorithms and other means, their behavior becomes more dependent upon the content of training data and other factors which cannot be anticipated by designers or operators.

Even when a single system is predictable, or even deterministic, when such systems interact with other systems, or in large masses or swarms, their collective behavior can become intrinsically unpredictable. This includes unpredictable interactions between known systems and adversarial systems whose operational behavior may be unknown.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

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 the systems of one's adversary and how an adversary will adapt to your system resulting in an inherently unpredictable act.

Candidate Recommendations

- Systems that exhibit intrinsically unpredictable behavior should be considered illegal and not deployed.
- Similarly, deploying systems with otherwise predictable behavior in situations or contexts in which the collective behavior of systems cannot be predicted should be avoided. In particular, deploying AWS swarms in which the emergent dynamics of the swarm have a significant impact on the actions of an individual AWS must be avoided.
- The predictability of weapons systems should be assessed with confidence levels with respect to specified contexts and circumstances of use. Systems should not be used outside of the contexts of use for which their operational behavior is understood and predictable. Engineers should explicitly examine their systems and inform their customers of their qualitative and quantitative confidence in the predictability of the actions of the autonomous functions of weapons systems in response to representative scenarios, specific contexts of use, and scope of operations.
- Commanders and operators should be trained to understand and assess confidence in the behavior of a system under specific contexts and scope of operations. They should maintain situational awareness of those contexts where weapons systems are deployed, and prevent those systems from being used outside the scope of operations for which their behavior is predictable.
- To ensure meaningful human control, operators should be able to query a system in real-time. Such a query should offer the evidence, explanation, and justification for critical determinations made by the systems, i.e., identification of a target, or key recommendations.
- Weapons systems with advance automation should also keep records and traces of critical functional and operational decisions that are made automatically. Such traces and records should be reviewable in instances where the behavior of the system was not as predicted.
- To the extent that systems contain adaptive or learning algorithms, any critical decision made by systems based upon those algorithms should be transparent and explainable by the designing engineers. Any data used for training and adaptation should be reviewed as to its integrity so as to ensure that learned functions can behave in reliably predictable ways.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Reframing Autonomous Weapons Systems

Further Resources

- International Committee for Robot Arms Control. "LAWS: Ten Problems for Global Security" (leaflet). April 10, 2015.
- Owens, D. "Figuring Forseeability." *Wake Forest Law Review* 44 (2009): 1277, 1281–1290.
- Scherer, M. "Who's to Blame (Part 5): A Deeper Look at Predicting the Actions of Autonomous Weapons." *Law and AI*, February 29, 2016.
- Arquilla, J., and D. Ronfeldt. *Swarming and the Future of Conflict*, Santa Monica, CA: RAND Corporation, 1997.
- Edwards, S. J. A. *Swarming and the Future of Warfare*, Santa Monica, CA: RAND Corporation, 2004.
- Rickli, J.-M. "Some Consideration of the Impact of Laws on International Security: Strategic Stability, Non-State Actors and Future Prospects." Meeting of Experts on Lethal Autonomous Weapons Systems Convention on Certain Conventional Weapons (CCW) United Nations Office Geneva, April 16, 2015.
- Scharre, P. *Robotics on the Battlefield Part II: The Coming Swarm*, Washington, DC: Center for a New American Security, 2014.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Economics and Humanitarian Issues

Autonomous and Intelligent systems (A/IS) provide unique and impactful opportunities in the humanitarian space. As disruptive technologies, they promise to upend multiple historical institutions and corresponding institutional relationships, offering opportunities to "re-intermediate" those settings with more humanitarian and equitably focused structures.

The value of A/IS is significantly associated with the generation of superior and unique insights, many of which could help to foster the accomplishment of humanitarian and development goals and to achieve positive socio-economic outcomes for both developed and developing economies. Among the opportunities for cooperation and collaboration at the intersection of A/IS and humanitarian and development issues are the following:

A/IS have been recognized as key enablers for achieving the goals of humanitarian relief, human rights, and the United Nations Sustainable Development Goals. This recognition provides the opportunity to demonstrate the positive and supportive role that A/IS can play in these critical, but perennially under-resourced and overlooked, areas.

A/IS are related to, but hold a unique place within, the larger "ICT for development" narrative. This intersection creates opportunities for A/IS to be applied in settings where commercial and development agendas meet, and to facilitate advances in the administration and impact assessment of development programs.

There is an ongoing narrative on affordable and universal access to communications networks and the Internet which invites consideration of how the implementations and fruits of A/IS will be made available to populations.

The narrative of "A/IS for the common good" is starting to present itself in various settings. Key elements framing this "common good" conversation relate to the need for it to be human-centered and include the need for accountability and to ensure that outcomes are fair and inclusive.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Economics and Humanitarian Issues

The scaling and use of A/IS represent a genuine opportunity to provide individuals and communities — be they rural, semi-urban, or cities — with greater autonomy and choice. A/IS will potentially disrupt all manner of economic, social, and political relationships and interactions. Those disruptions will provide a historical opportunity to re-establish those settings so that they are reflective of more updated and sustainable notions of autonomy and choice.

Many of the debates surrounding A/IS 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, excessive imbalances, classism, and general non-acceptance of these technologies.

In the absence of that comprehensive treatment, A/IS policy issues will be addressed piecemeal by different jurisdictions and in different sectors. In that context of "distributed policy making," a patchwork of policies and initiatives is the likely result, dissipating potential impact. However, some measure of "policy interoperability" can still be served if there is a common framing or policy generation process for analysis that can be shared across jurisdictions and/or sectors.

The use of A/IS in support of the pragmatic outcomes noted above is best framed within four key domains that comprise the following four sections: **economics, privacy and safety, education,** and **equal availability**. Each of these contexts presents unique challenges, attention to which can inform the trustworthy use of A/IS for the common good.

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 1 — Economics

While the increase of A/IS and its positive uses in society are undeniable, the financial gains from these technologies may favor certain sectors, and are not evenly distributed throughout populations where it is created or deployed. Likewise, while A/IS automation of certain human tasks may be beneficial by supplanting arduous jobs, how employment in aggregate for specific populations and job verticals will be affected by A/IS needs to be addressed.

Issue:

A/IS should contribute to achieving the UN Sustainable Development Goals.

Background

The contribution of A/IS to human and sustainable development in developing countries, and in particular extreme poverty eradication, is inherently connected with its contribution to human well-being in the developed world. In a globalizing society, one part of the world has a direct impact on another. With a growing level of interdependence between communities, the challenges and opportunities are truly global. Climate change, poverty, globalization, and technology are closely interconnected. Ethical commitment should entail a sense of global citizenship and of responsibility as members of humanity.

Beyond considering the humanitarian role of A/IS, there is a pressing need to address how these technologies can contribute to achieving the UN Sustainable Development Goals that concern eradicating poverty, illiteracy, gender and ethnic inequality, and combating the impact of climate change.

The inequality gap between the developed and the developing nations is disturbingly wide. With the introduction of hi-tech, the world had witnessed a considerable increase in the existing gap as the new market is dominated by products and services from this new sector. One of the factors contributing to this is the nature of the tech economy and its tendency to concentrate wealth in the hands of few. The tech economy is also susceptible to corporate aggregation.

We need to answer questions such as "How will developing nations implement A/IS via existing resources? Do the economics of developing nations allow for A/IS implementation? What should be the role of the public and the private sectors and society in designing, developing, implementing, and controlling A/IS? How can people without technical expertise maintain these systems?"

The risk of unemployment for developing countries is more serious than for developed countries. The industry of most developing

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Economics and Humanitarian Issues

countries is labor intensive. While labor may be cheap(er) in developing economies, the ripple effects will be felt much more than in the developed economies as more and more jobs will be gradually replaced along with the development of robots or A/IS.

As an example, in the manufacturing industry, lots of products such as mobile phones and clothes are designed in developed countries, but made in developing countries. Thus, it is not difficult to predict that the developing countries will be at greater risk of unemployment than developed countries if those manufacturing tasks can be replaced by machines. The challenge of unemployment is even bigger for developing countries than for developed countries, which can exacerbate the economic and power-structure differences between and within developed and developing nations.

Candidate Recommendations

The current panorama of applications of A/IS in sectors crucial to the UN Sustainable Development Goals should be studied, and the strengths, weaknesses, and potential of some of the most significant recent applications drawn from these sectors should be analyzed. *Specific areas to consider include:*

- Taking appropriate action to mitigate the gap. The private sector should integrate CSR (corporate social responsibility) at the core of development and marketing strategies and operations. Mitigating the social problems of technology development should be a special focus of responsible companies using A/IS.
- Developing mechanisms for increasing transparency of power structures and justly sharing the economic and knowledge acquisition benefits of robotics/A/IS.
- Facilitating robotics/A/IS research and development in developing nations.
- Empowering the education sector with advanced courses on A/IS is the first step toward creating a nation that can handle the new economic and power shifts.
- Investing in technology transfer will help developing nations reduce the gap.
- Adapting legal and policy frameworks which will help to favor equitable distribution of wealth, empowering competent international organizations to favor a minimally viable competition level on the A/IS markets to avoid detrimental monopolistic situations.
- Identifying A/IS technologies relevant to the UN Sustainable Development Goals such as big data for development (agriculture, medical tele-diagnosis), geographic information systems (disaster prevention, emergency planning), and control systems (naturalizing intelligent cities through energy and traffic control, management of urban agriculture).
- Developing guidelines and recommendations for the nurturing and implementation of these technologies in developing countries.
- Documenting and disseminating successful examples of good practice, and evaluations and conclusions of experiences.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Economics and Humanitarian Issues

- Developing appropriate impact indices for the evaluation of A/IS technological interventions in developing countries from multiple perspectives.

Further Resources

- United Nations. "Sustainable Development Goals: 17 Goals to Transform Our World." September 25, 2015.

Issue:

It is unclear how developing nations can best implement A/IS via existing resources.

Background

Do the economics of developing nations allow for A/IS implementation? How can people without technical expertise maintain design specifications and procure these systems? The potential use of A/IS to create sustainable economic growth for LMICs (lower and middle income countries) is uniquely powerful. If A/IS capacity and governance problems are addressed, LMICs will have the ability to use A/IS to transform their economies and leapfrog into a new era of inclusive growth if a clear path for development is provided. Particular attention, however, should be paid to ensure that the use of A/IS for the common good — especially in the context of LMICs — does not reinforce existing socio-economic inequities.

Candidate Recommendations

- Develop mechanisms for increasing transparency of power structures and justly sharing the economic and knowledge acquisition benefits of A/IS.
- Facilitate A/IS research and development in developing nations. Ensure that representatives of developing nations are involved.
- Along with the use of A/IS, discussions related to identity, platforms, and blockchain are needed to ensure that all of the core enabling technologies are designed to meet the needs of LMICs.

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.
- Bloom, N. "Corporations in the Age of Inequality." *Harvard Business Review,* April 21, 2017.
- Dahlman, C. *Technology, Globalization, and Competitiveness: Challenges for Developing Countries. Industrialization in the 21st Century.* New York: United Nations, 2006.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Economics and Humanitarian Issues

- 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, U.K.: 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 Rich and the Rest." *The Economist.* October 13, 2012.
- "Wealth Without Workers, Workers Without Wealth." *The Economist*. October 4, 2014.
- World Bank. "Global Economic Prospects 2008: Technology Diffusion in the Developing World." Washington, DC: World Bank, 2008.

Issue:

The complexities of employment are being neglected regarding A/IS.

Background

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

Candidate Recommendations

It is important to focus the analysis on how employment structures 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.

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. *Specifics to accomplish this include:*

- Create an international, independent agency which can properly disseminate objective statistics and inform media as well as the general public about the impact of robotics and A/IS on jobs and growth.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Economics and Humanitarian Issues

- Consider both product and process innovation and look at it from a global perspective as a way to understand properly the global impact of A/IS on employment (refer to Pianta, 2009 and Vivarelli 2007).
- Focus the analysis on how employment structures will be changed by A/IS rather than 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 (refer to Bonin et al. 2015 and RockEU, 2016).
- Make sure workers can improve their adaptability to fast technological changes by providing them adequate training programs. Those training programs could be available to any worker with a special attention to low-skilled workforce members. 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.
- Ensure that not only the worker whose job is concerned benefits from training programs, but also any employee in the company so everyone has the chance to be up to speed with technical changes, even if one's job is not immediately concerned (not only *reaction* but also *prevention*). Thus it should be the responsibility of every company to increase its investment in the internal training of its workforce based on the profitability gains realized thanks to automation. The public side could facilitate such initiatives with co-investment in the training programs through tax incentives.

Further Resources

- RockEU. "Robotics Coordination Action for Europe Report on Robotics and Employment," Deliverable D3.4.1, June 30, 2016.
- International Federation of Robotics. "The Impact of Robots on Productivity, Employment and Jobs," A positioning paper by the International Federation of Robotics, April 2017.
- 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.

Issue:

Automation is often viewed only within market contexts.

Background

A/IS 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 A/IS. This impact will diffuse through the global society.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Economics and Humanitarian Issues

Candidate Recommendation

To understand the impact of A/IS on society, it is necessary to consider both product and process innovation as well as wider implications from a global perspective.

Further Resources

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

Issue:

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

Background

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

The wave of automation caused by the A/IS revolution will displace a very large amount of jobs across domains and value chains. The U.S. "automated vehicle" case study analyzed in the White House 2016 report *Artificial Intelligence, Automation, and the Economy* is emblematic of what's at stake: 2.2 to 3.1 million existing part- and full-time U.S. jobs are exposed over the next two decades, although the timeline remains uncertain. In particular, between 1.3 and 1.7 million heavy truck drivers are threatened. And this is not trivial, for the profession has symbolized in the collective imagination the manifestation of the American dream of empowerment, liberty, and social ascension whereby less-educated people could make it into the middle class.

The automation wave calls at least for higher investment and probably the need to reinvent active labor market programs in the coming decades. Such investment should logically be funded by fiscal policies targeting the capital. The 2016 White House report gave an interesting order of magnitude applied to the case of the United States: "increasing funding for job training in the U.S. by six-fold — which would match spending as a percentage of GDP to Germany, but still leave the U.S. far behind other European countries — would enable retraining of an additional 2.5 million people per year."

A/IS and other digital technologies offer real potential to innovate new approaches to job-search assistance, placement, and hiring processes in the age of personalized services. The efficiency of matching labor supply and demand can be tremendously enhanced by the rise of multi-sided platforms and predictive analytics. The case of platforms, such as LinkedIn for instance with its 470 million registered users, is interesting as an evolution in hiring practices.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Economics and Humanitarian Issues

Tailored counseling and integrated re-training programs also represent promising grounds for innovation.

This, however, will not be enough. A lot will have to be done to create fair and effective life-long skill development/training infrastructure and mechanisms capable of empowering millions of people to viably transition jobs, sectors, and potentially geographies. A lot will also have to be done to address differential geographic impacts which exacerbate income and wealth disparities. Effectively enabling the workforce to be more mobile — physically, legally, and virtually — will be crucial. This implies systemic policy approaches which encompass housing, transportation, licensing, taxes, and, crucially in the age of A/IS, broadband access, especially in rural areas.

Candidate Recommendations

- To cope with the technological pace and ensuing progress of A/IS, it will be necessary for workers to 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.
- To lay solid foundations for the profound transformation outlined above, more research in at least three complementary areas is needed:
    - First, to devise mechanisms of dynamic mapping of tasks and occupations at risks of automation and associated employment volumes. This mapping of the workforce supply is needed at the macro, but also crucially at the micro, levels where labor market programs are deployed;
    - Integrated with that, more granular and dynamic mapping of the future jobs/tasks, workplace-structures, associated work-habits, and skill-base spurred by the A/IS revolution are also needed. This mapping of the demand side will be key to innovate, align, and synchronize skill development and training programs with future requirements.
    - More policy research on the dynamics of professional transitions in different labor market conditions is required.
- To maximize intended impact, create necessary space for trial-and-error strategies, and to scale up solutions that work, implement robust, data-driven evidence-based approaches. These approaches should be based on experiments and centered on outcomes in terms of employment but also in terms of earnings. New forms of people-

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Economics and Humanitarian Issues

public-private partnerships involving civil society as well as new outcome-oriented financial mechanisms (social impact bonds, for instance) that help scale up successful innovations should also be considered.

- The next generation of highly qualified personnel should be ready to close skills gaps and develop future workforces. New programs should be offered possibly earlier than high school, to increase access to employment in the future.

Further Resources

- Executive Office of the President. Artificial Intelligence, Automation, and the Economy. December 20, 2016.
- Kilcarr, S. "Defining the American Dream for Trucking ... and the Nation, Too," *FleetOwner*, April 26, 2016.
- OECD, "Labor Market Programs: Expenditure and Participants," *OECD Employment and Labor Market Statistics* (database), 2016.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 2 — Privacy and Safety

The growing volumes of private sector data (mobile phones, financial transactions, retail, logistics) hold unique promise in developing more robust and actionable disease-monitoring systems that can be empowered by A/IS. However, concerns related to privacy, the ability of individuals to opt out, the cross-border nature of data flows, and the political and commercial power dynamics of this data are the key factors to consider in how to most equitably shape this domain.

Issue:

There is a lack of access and understanding regarding personal information.

Background

How to handle privacy and safety issues, especially as they apply to data in humanitarian and development contexts? 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 the ambiguities of data "ownership," privacy, and how their digital access generates personal data by-products used by third parties.

According to the 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% of the developing world population to 45% over the period. In Sub-Saharan Africa, just 17% of the population were mobile Internet subscribers in 2013, but penetration is forecast to increase to 37% 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.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Economics and Humanitarian Issues

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.

Candidate Recommendation

Frameworks such as Privacy by Design can guide the process of identifying appropriate system and software requirements in early stages of design. Such frameworks also encourage proactive examination of harms and risks, seek to engage the data subject (e.g., consumer, user, stakeholders) in the design of the software, and recommend best practices and regulatory requirements (such as data minimization, accountability, transparency, options such as opt-in, opt-out, encryption) to be embedded into the system. *In addition:*

- Best practices such as Privacy Impact Assessments will assist with identification of data misuse cases at early stages of system/software design.
- Improving digital literacy of citizens should be a high priority for the government and other organizations.
- Governments should enforce transparency related to data collection, data ownership, data stewardship, and data usage and disclosure.
- Organizations should be held accountable for data misuse, financial loss, and harm to the reputation of the data object if data is mishandled. This requires that organizations have appropriate policies and agreements in place, that terms and conditions of the agreements are clearly communicated with the data object and that data misuse cases and legitimate use cases are well-defined in advance.

Further Resources

- For more on responsible data use, please see the section "Personal Data and Individual Access Control."
- For more on responsible data use, see the Responsible 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.
- van Rest J., D. Boonstra, M. Everts, M. van Rijn, R. van Paassen. "Designing Privacy-by-Design," in *Privacy Technologies and Policy,* edited by B. Preneel, and D. Ikonomou. Lecture Notes in Computer Science, vol. 8319. Berlin, Germany: Springer, 2012.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 3 — Education

It is essential to increase the awareness, critical understanding, and attitudinal values of undergraduate and postgraduate students related to sustainable human development and its relationship with A/IS, so that they are prepared to assume their responsibilities in the solution of the current global social crises. Current and future leaders should be educated in macro-ethics and not only in micro-ethics.

Shared narratives, generated by awareness, education, and standard evaluative models are the best pathway to generating the global support necessary to meet these challenges. Programs fostering awareness, education, and analytical and governance models should address the opportunities and risks of A/IS in development contexts.

Issue:

How best to incorporate the "global dimension of engineering" approach in undergraduate and postgraduate education in A/IS.

Background

A/IS presents a unique opportunity for narrative and policy construction in educational institutions. Where norms exist, they are taught in schools. Thus, physics majors learn the "standard" theories and equations of physics. The same is true in other disciplines. However, where standards are either absent or in the process of development in a sector, what is most appropriately included in undergraduate and graduate curriculum is less clear. That is the case for a number of areas in the digital world, including A/IS. Thus, educators and other parties involved in curriculum development should consider the opportunity to craft curricula that will make their students aware of this absence of standards, and also encourage the exploration of various practices as candidates for "best practices" and their possible further elevation to standards in AI technology and policy.

Candidate Recommendations

The understanding of the global dimension of engineering practice should be embedded in A/IS curricula. *Specifically:*

- Curriculum/core competencies should be defined and preparation of course-material repositories and choice of the most adequate pedagogical approaches should be established.
- The potential of A/IS applications should be emphasized in undergraduate and graduate programs specifically aimed at engineering in international development and humanitarian relief contexts as well as in the training programs preparing technical professionals for work in the international development and humanitarian sectors.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Economics and Humanitarian Issues

- Increased awareness on the opportunities and risks faced by Lower Middle Income Countries (LMICs) in the use of A/IS for sustainable development and humanitarian purposes is critical. Ignoring these opportunities and risks will further divide the opportunities for development across the globe. A/IS presents an opportunity to potentially reduce these differentials that ultimately strain social fabric and economic systems.

Further Resource

- Global Dimension in Engineering Education Project (GDEE).

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 4 — Equal Availability

For A/IS to be adopted in an atmosphere of trust and safety, greater efforts must be undertaken to increase availability of these resources.

Issue:

AI and autonomous technologies are not equally available worldwide.

Background

Equitable distribution of the benefits of A/IS technology worldwide should be prioritized. Training, education, and opportunities in A/IS worldwide should be provided particularly with respect to underdeveloped nations.

Candidate Recommendations

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

1. Engage in discussions regarding effective A/IS education and training.
2. Encourage global standardization/ harmonization and open source software for A/IS.
3. Promote distribution of knowledge and wealth generated by the latest A/IS, including formal financial mechanisms (such as taxation or donations to effect such equity worldwide).
4. International organizations, government bodies, universities, and research institutes should promote research into A/IS technologies that are readily available in developing countries, for example, mobile lightweight A/IS applications (taking advantage of the widespread use of increasingly affordable Internet-enabled phones in developing contexts) and culture-aware systems.
5. National and international development cooperation agencies and NGOs should draw attention to the potential role of A/IS in human and sustainable development.

Further Resources

- Hazeltine, B., and C. Bull. *Appropriate Technology: Tools, Choices, and Implications.* New York: Academic Press, 1999.
- Akubue, A. "Appropriate Technology for Socioeconomic Development in Third World Countries." *The Journal of Technology Studies* 26, no. 1 (2000): 33–43.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

The first edition of the law section for *Ethically Aligned Design* noted that the early stages in development of autonomous and intelligent systems (A/IS) have given rise to many complex ethical problems that translate directly and indirectly into discrete legal challenges. That is, of course, what the rule of law often intends to answer — how we should behave as a society when faced with difficult ethical decisions — and it should come as no surprise that the legal implications of A/IS continue to unfold as we witness the forms of its expression and use expand.

To consider the ongoing creep of A/IS ethical issues into the legal realm, one need look no further than the first section of this document: *Legal Status*. This section addresses what legal status should A/IS be granted and was not a topic in the original edition. That is to say, in just one revision of this paper, we felt the need to address the question of how A/IS should be labeled in the courts' eyes: a product that can be bought and sold? A domesticated animal with more rights than a simple piece of property, but less than a human? A person? Something new?

Our conclusion to that question is that A/IS are not yet deserving of any kind of "personhood" — yet the very fact that the question of whether A/IS could, or should, be granted such status demonstrates the rate at which the technology and the related legal and ethical questions are growing and provide two universal principles echoed throughout this document:

The development, design, and distribution of A/IS should fully comply with all applicable international and domestic law.

There is much work to be done: the legal and academic community must increase engagement in this rapidly developing field from its members.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

Concerns and recommendations fall into four main areas:

1. Legal Status of A/IS

2. Governmental Use of A/IS: Transparency and Individual Rights

3. Legal Accountability for Harm Caused by A/IS

4. Transparency, Accountability, and Verifiability in A/IS

While much debate continues to surround A/IS, its development, and use, these questions must be addressed *before* the proliferation of A/IS passes some kind of tipping point. The authors hope this paper will inform the legislative process and inspire more members of the legal community to become involved *now*.

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

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

Section 1 — Legal Status of A/IS

There has been much discussion about how to legally regulate A/IS-related technologies, and the appropriate legal treatment of systems that deploy these technologies. Lawmakers today are wrestling with the issue of what status to apply to A/IS. Legal "personhood" (as is applied to humans and certain types of human organizations) is one possible option for framing such legal treatment, and the implications of granting that status to A/IS applications raises issues that have implications in multiple domains of human interaction beyond technical issues.

Issue:

What type of legal status (or other legal analytical framework) is appropriate for application to A/IS, given the legal issues raised by deployment of such technologies?

Background

The convergence of A/IS and robotics technologies has led to the development of systems and devices with attributes that resemble those of human beings in terms of their autonomy, ability to perform intellectual tasks and, in the case of some robots, their physical appearance. As some types of A/IS begin to display characteristics that resemble those of human actors, some governmental entities and private commentators have concluded that it is time to examine how legal regimes should categorize and treat various types of A/IS. These entities have posited questions such as, "Should the law treat such systems as legal 'persons,' with all the rights and responsibilities that personhood entails?" Such status seems initially remarkable until consideration is given to the long-standing legal personhood status granted to corporations, governmental entities, and the like — none of which are human even though they are run by humans.

Alternatively, many entities have asked, should some A/IS be treated as mere products and tools of their human developers and users? Perhaps A/IS are something entirely without precedent, raising the question of whether one or more types of A/IS might be assigned an intermediate — and perhaps novel — type of legal status?

Clarifying the legal status of A/IS in one or more jurisdictions is essential in removing the uncertainty associated with the obligations and expectations for organization and operation of these systems. Clarification along these lines will encourage more certain development and deployment of A/IS and will help clarify lines of legal responsibility and liability when A/IS cause harm. Recognizing A/IS as independent "legal persons" would, for example, limit or eliminate

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

some human responsibility for subsequent "decisions" made by such A/IS (for example under a theory of "intervening causation" — akin to the "relief" from responsibility of a hammer manufacturer when a burglar uses a hammer to break the window of a house), thus potentially reducing the incentives for designers, developers, and users of A/IS to ensure their safety. In this example, legal issues that are applied in similar "chain of causation" settings (such as "foreseeability," "complicity," "reasonable care," "strict liability" for unreasonably dangerous goods, and other precedential notions) will factor into the design process. Different jurisdictions may reach different conclusions about the nature of such causation chains, inviting future creative legal planners to consider how and where to pursue design, development, and deployment of future A/IS in order to receive the most beneficial legal treatment.

The issue of the legal status of A/IS thus intertwines with broader legal questions regarding how to ensure accountability and assign and allocate liability when A/IS cause harm. The question of legal personhood for A/IS also interacts with broader ethical questions on the extent to which A/IS should be treated as moral agents independent from their human designers and operators, and whether recognition of A/IS personhood would enhance or detract from the purposes for which humans created the A/IS in the first place.

A/IS are at an early stage of development where it is premature to assert a single particular legal status or presumption for application in the many forms and settings in which those systems are deployed. This uncertainty, coupled with the multiple legal jurisdictions in which A/IS are being deployed (each of which, as a sovereign, can regulate A/IS as it sees fit) suggests that there are multiple general frameworks through which to consider A/IS legal status. Below are some examples.

Candidate Recommendations

1. While conferring legal personhood on A/IS might bring some economic benefits, the technology has not yet developed to the point where it would be legally or morally appropriate to generally accord A/IS the rights and responsibilities inherent in the legal definition of personhood, as it is defined today. Therefore, even absent the consideration of any negative ramifications from personhood status, it would be unwise to accord such status to A/IS at this time. A/IS should therefore remain to be subject to the applicable regimes of property law.

2. Government and industry stakeholders alike should identify the types of decisions and operations that should never be delegated to A/IS, and adopt rules and standards that ensure effective human control over those decisions. Modern legal systems already address a number of other situations that could serve as appropriate analogues for the legal status of A/IS and how to allocate legal responsibility for harm caused by A/IS. These legal analogues may include the treatment of pets, livestock, wild animals, employees, and other "agents" of persons and corporations. Governments and courts should review

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

these various potential legal models and assess whether they could serve as a proper basis for assigning and apportioning legal rights and responsibilities with respect to the deployment and use of A/IS.

3. In addition, governments should scrutinize existing laws — especially those governing business organizations — for mechanisms that could have the practical effect of allowing A/IS to have legal autonomy. If ambiguities or loopholes in the law could create a legal method for recognizing A/IS personhood, the government should review and, if appropriate, amend the pertinent laws.

4. Manufacturers and operators should gain an understanding of how each jurisdiction would categorize a given A/IS and how each jurisdiction would treat harm caused by the system. Manufacturers and operators should be required to comply with the applicable laws of all jurisdictions in which that system could operate. In addition, manufacturers and operators should be aware of standards of performance and measurement promulgated by standards development organization and agencies.

5. As A/IS become more sophisticated, governments should reassess the issue of legal status for these systems. In considering whether to accord legal protections, rights, and responsibilities to A/IS, governments should exercise utmost caution. Governments and decision-makers at every level must work closely with regulators, representatives of civil society, industry actors, and other stakeholders to ensure that the interest of humanity — and not the interests of the autonomous systems themselves — remains the guiding principle.

Further Resources

- Bayern, S. "The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems." *Stanford Technology Law Review* 19, no. 1 (2015): 93–112.
    - Bayern, S. et al., "Company Law and Autonomous Systems: A Blueprint for Lawyers, Entrepreneurs, and Regulators." *Hastings Science and Technology Law Journal* 9, no. 2 (2017): 135–162.
- Bhattacharyya, D. "Being, River: The Law, the Person and the Unthinkable." *Humanities and Social Sciences Online,* April 26, 2017.
- Calverley, D. J. "Android Science and Animal Rights, Does an Analogy Exist?" *Connection Science* 18, no. 4 (2006): 403–417.
- Calverley, D. J. "Imagining a Non-Biological Machine as a Legal Person." *AI & Society* 22 (2008): 403–417.
- European Parliament Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics.
- Zyga, L. "Incident of Drunk Man Kicking Humanoid Robot Raises Legal Questions," *Techxplore,* October 2, 2015.
- LoPucki, L. M. "Algorithmic Entities." *Washington University Law Review* 95 (forthcoming 2017).

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

- Scherer, M. "Digital Analogues." *Imaginary Papers,* June 8, 2016.
- Scherer, M. "Is Legal Personhood for AI Already Possible Under Current United States Laws?" *Law and AI,* May 14, 2017.
- Solum, L. B. "Legal Personhood for Artificial Intelligences." *North Carolina Law Review* 70, no. 4 (1992): 1231–1287.
- Weaver, J. F. Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws. Santa Barbara, CA: Praeger, 2013.
- European Parliament. European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics. February 16, 2017.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

Section 2 — Governmental Use of A/IS: Transparency and Individual Rights

Surveillance of populations by governments and the disruption of free elections will become ever easier as we deploy A/IS. How should we manage these systems to ensure that they act for the good of society?

Issue:

International, national, and local governments are using A/IS. How can we ensure the A/IS that governments employ do not infringe on citizens' 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 A/IS make important decisions about individuals.

Candidate Recommendations

1. Government stakeholders should identify the types of decisions and operations that should never be delegated to A/IS, such as when to use lethal force, and adopt rules and standards that ensure effective human control over those decisions.

2. Governments should not employ A/IS 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 A/IS and an ability to explain its logical reasoning must be required. Given the current abilities of A/IS, under no circumstances should court decisions be made by such systems alone. Parties, their lawyers, and courts must have reasonable access to all data and information generated and used by A/IS technologies employed by governments and other state authorities.

3. A/IS 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

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

inspection, then alternative ways must be available to uphold the values of transparency. Such systems should be subject to risk assessments and rigorous testing.

4. Individuals should be provided a forum to make a case for extenuating circumstances that the A/IS 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.

5. Automated systems should generate audit trails recording the facts and law supporting decisions and such systems should be amenable to third-party verification to show that the trails reflect what the system in fact did. Audit trails should include a 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. Providers of A/IS, or providers of solutions or services that substantially incorporate such systems, should make available statistically sound evaluation protocols through which they measure, quality assure, and substantiate their claims of performance, for example, relying where available on protocols and standards developed by the National Institute of Standards and Technology (NIST) or other standard-setting bodies.

6. Investor list(s), developers, and promoters of any given A/IS being developed should be required by law to be made public when the A/IS are used for governmental purposes. There should also be transparency of the specific ethical values promoted by the designer, and *how* they were embedded in the system. Transparency should also apply to the input data selection process.

Further Resources

- Schwartz, P. "Data Processing and Government Administration: The Failure of the American Legal Response to the Computer." *Hastings Law Journal* 43 (1991): 1321–1389.

- Citron, D. K. "Technological Due Process." Washington *University Law Review* 85 (2007): 1249–1313.

- Citron, D. K. "Open Code Governance." *University of Chicago Legal Forum* 2008, no. 1 (2008): 355–387.

- Crawford, K., and J. Schultz. "Big Data and Due Process: Toward a Framework to Address Predictive Privacy Harms." *Boston College Law Review* 55, no. 1 (2014): 93–128.

- Pasquale, F. Black Box Society. Cambridge, MA: Harvard University Press, 2014.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

- Bamberger, K. "Technologies of Compliance: Risk and Regulation in the Digital Age." *Texas Law Review* 88, no. 4 (2010): 669–739.
- Kroll, J. Accountable Algorithms. Princeton, NJ: Princeton University Press, 2015.
- Desai, D., and J. A. Kroll. "Trust But Verify: A Guide to Algorithms and the Law." *Harvard Journal of Law and Technology,* forthcoming.
- ICRC. "Views of International Committee of Red Cross (ICRC) on Autonomous Weapon System." April 11, 2016.
- Rainie, L., J. Andesrson, and J. Albright. "The Future of Free Speech, Trolls, Anonymity and Fake News Online." Pew Research Center, March 29, 2017.
- Marwick, A., "Are There Limits to Online Free Speech?" *Data & Society: Points,* January 5, 2017.
- Neier, A., "Talking Trash: What's More Important, Human Dignity or Freedom of Speech?" *Columbia Journalism Review*, September/October 2012.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 3 — Legal Accountability for Harm Caused by A/IS

As A/IS becomes more prevalent while also potentially becoming more removed from the human developer/manufacturer, what is the correct approach to ensure legal accountability for harms caused by A/IS?

Issue:

How can A/IS 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 A/IS become more sophisticated and ubiquitous, that will increasingly be less true. The A/IS 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 A/IS 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 A/IS is not abused by businesses and entities employing the A/IS to avoid liability or payment of damages. Governments should consider adopting regulations requiring insurance or other guarantees of financial responsibility so that victims can recover damages for harm that A/IS cause.

3. Companies that use and manufacture A/IS should be required to establish written policies governing how the A/IS should be used, including the real-world applications for such AI, any preconditions for its effective use, who is qualified to use it, what training is required for operators, how to measure the performance of the A/IS, and what operators and other people can expect from the A/IS. This will help to give the human operators and beneficiaries an accurate idea of what to expect from the A/IS, while also protecting the companies that make the A/IS from future litigation.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

4. Because the person who activates the A/IS will not always be the person who manages or oversees the A/IS while it operates, states should avoid adopting universal rules that assign legal responsibility and liability to the person who "turns on" the A/IS. For example, liability may attach to the manufacturers or to the person who directs, monitors, and controls the A/IS's operations, or has the responsibility to do so.

6. For the avoidance of repeated or future harm, companies that use and manufacture A/IS should consider the importance of continued algorithm maintenance. Maintenance is an essential aspect of design. Design does not end with deployment. Thus, there should be a clear legal requirement of (1) due diligence, and (2) sufficient investment in algorithm maintenance on the part of companies that use and manufacture A/IS that includes monitoring of outcomes, complaint mechanism, inspection, correction, and replacement of harm-inducing algorithm, if warranted. Companies should be prohibited from contractually delegating this responsibility to unsophisticated end-users.

7. Promote international harmonization of national legislations related to liability in the context of A/IS design and operation (through bi- or multilateral agreements) to enhance interoperability, and facilitate transnational dispute resolution.

8. Courts weighing A/IS litigation cases based on some form of injury should adopt a similar scheme to that of product liability litigation, wherein companies are not penalized or held responsible for installing post-harm fixes on their products designed to make the product safer. In other words, because courts have recognized that it is good public policy to encourage companies to fix dangerous design flaws, retroactively fixing a design flaw that has caused injury is not considered an admission or a sign of culpability. The same approach should be used in A/IS litigation.

Further Resources

- Allan, T., and R. Widdison. "Can Computers Make Contracts?" *Harvard Journal of Law and Technology* 9 (1996): 25–52.
- Asaro, P. M. "The Liability Problem for Autonomous Artificial Agents." Palo Alto, CA: Association for the Advancement of Artificial Intelligence, 2015.
- Chopra, S., and L. F. White. A Legal Theory for Autonomous Artificial Agents. Ann Arbor, MI: University of Michigan Press, 2011.
- Colonna, K, "Autonomous Cars and Tort Liability." *Case Western Journal of Law, Technology & The Internet* 4 no. 4 (2012): 81–130.
- Field, C. "South Korean Robot Ethics Charter 2012." PhD thesis (part), Sydney, Aus.: University of Technology, 2010.
- Grossman, M. R., and G. V. Cormack. "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review." *Richmond Journal of Law and Technology* 17, no. 3 (2011): 1–48.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

- Kalra, N., J. Anderson, and M. Wachs. "Liability and Regulation of Autonomous Vehicle Technologies." State of California Department of Transportation Technical Report. Berkeley, CA: Institute of Transportation Studies, University of California, 2009.
- Krakow, C. E. A. "Liability for Distributed Artificial Intelligences." *Berkeley Technology Law Journal* 11, no. 1 (1996): 147–204.
- Rivard, M. D. "Toward a General Theory of Constitutional Personhood: A Theory of Constitutional Personhood for Transgenic Humanoid Species." *UCLA Law Review* 39, no. 5 (1992): 1425–1510.
- Scherer, M., "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." *Harvard Journal of Law and Technology* 29, no. 2 (2016): 353–400.
- Tobin, R., and E. Schoeman. "The New Zealand Accident Compensation Scheme: The Statutory Bar and the Conflict of Laws." *The American Journal of Comparative Law* 53, no. 2 (2005): 493–514.
- Wachter, S., B. Mittelstadt, and L. Floridi. "Transparent, Explainable, and Accountable AI for Robotics." *Science Robotics* 2, no. 6 (May 31, 2017). DOI: 10.1126/scirobotics.aan6080.
- Weaver, J. F.. Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws. Santa Barbara, CA: Praeger, 2013.
- Weiss, A. "Validation of an Evaluation Framework for Human-Robot Interaction. The Impact of Usability, Social Acceptance, User Experience, and Societal Impact on Collaboration with Humanoid Robots." PhD thesis, University of Salzburg, 2010.
- Wooldridge, M., and N. R. Jennings. "Intelligent Agents: Theory and Practice." *The Knowledge Engineering Review* no. 2 (1995): 115–152.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 4 — Transparency, Accountability, and Verifiability in A/IS

Transparency around A/IS is a difficult issue because it impinges on the differing needs of developers for trade secrets and users to be able to understand the technology to guard against problems occurring with it, and to hold accountable the correct entity in the event of a system failure.

Issue:

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

Background

Decision-making algorithms can be designed for various purposes, and the applications are wide-ranging for both the public and the private sectors. We must assume that virtually every decision that we make as humans can be mediated or replaced by an algorithm. Therefore, we cannot overestimate both the current and future role of A/IS across different sectors. Algorithms and automated decision-making (e.g., resume/cv screening during job applications) have the potential to be more fair, and less biased than humans, provided that the systems are designed well. This requires, in particular, that effective preventative measures are put in place to avoid an algorithm-based information and/or value bias.

At the same time, most users of A/IS will not be aware of the sources, scale, varying levels of accuracy, intended purposes, and significance of uncertainty in the operations of A/IS, or that they are interacting with A/IS in the first place. The sources of data used to perform these tasks are also often unclear. Furthermore, users might not foresee the inferences that can be made about them or the consequences when A/IS are used. The proliferation of A/IS will result in an increase in the number of systems that rely on machine learning and other developmental systems whose actions are not pre-programmed, and that may not produce logs or a record of how the system reached its current state.

These systems are often opaque (frequently referred to as "black boxes") and create difficulties for everyone, from the engineer, to the lawyer in court, to the online shopper, to the social media user. The result is an abundance of ethical issues of ultimate accountability.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

Candidate Recommendations

1. Given that many of the desired design specifications regarding accountability and verifiability are not technologically possible at this time, for now, this is an ethical issue that is best addressed by disclosure. If users are aware that they are interacting with an A/IS in the first place, and know exactly what information is being transferred to it, they will be better suited to tailor their inputs. A government-approved labeling system like the skull and crossbones found on household cleaning supplies that contain poisonous compounds could be used for this purpose to improve the chances that users are aware when they are interacting with A/IS.

2. Designers and manufacturers must remain accountable for the risks or externalities their systems cause. This is a balancing act since the level of risk that is acceptably mitigated through disclosure is not always clear. Recommending specific levels (whether a manufacturer of A/IS acts responsibly, or whether there is enough disclosure, or whether total disclosure would even be enough to mitigate the risk to users) is often a fact-specific discussion that doesn't suit itself well to broad rules.

3. There is a demand for algorithmic operation transparency. Although it is acknowledged this cannot be done currently, A/IS 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 to the extent possible sources of uncertainty, and state any assumptions relied upon.

4. A/IS should be programmed so that, under certain high risk situations where human decision-making is involved, they proactively inform users of uncertainty even when not asked.

5. With any significant potential risk of economic or physical harm, designers should conspicuously and adequately warn users of the risk and provide a greater scope of proactive disclosure to the user. Designers should remain mindful that some risks cannot be adequately warned against and should be avoided entirely.

6. To reduce the risk of A/IS that are unreasonably dangerous or that violate the law from being marketed and produced, we recommend lawmakers provide whistleblower incentives and protections. As in many industries, insiders may often be the first to know that the A/IS are acting illegally or dangerously. A well-crafted law to protect whistleblowers and allow a public interest cause of action would improve accountability and aid in prevention of intentional, reckless, or negligent misuses of A/IS.

7. Government and industry groups should consider establishing standards that require A/IS to create logs (or other means of verification of their decision-making process) regarding key aspects of their operations and store those logs for a specified period of time. Designers should leverage current computer science regarding accountability and verifiability for code. New verification techniques may need to be developed

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

to overcome the technical challenges in verifiability and auditability of A/IS operations; A/IS oversight systems ("A/IS guardians") or methods such as Quantitative Input Influence ("QII") measures could facilitate this process. Making sure, *ex ante*, that such information is, or can be made, available will also provide a higher degree of trust in verifiability and a sense of transparency in A/IS operations.

8. In Europe, the discussion on the so called "right to explanation" when automated decision-making occurs is important to address. Even though it is not yet guaranteed in Europe, future jurisprudence or Member State laws could grant individuals the right to ask for an explanation when a solely automated decision (e.g., refusal of an online credit application or e-recruiting practices) is being made about them that has legal or other significant effects. Such a right could provide a mechanism to increase the transparency and accountability of A/IS, and should therefore be seriously considered. In addition, other accountability enhancing tools such as ethical audits or certification schemes for algorithms should be explored. In addition, users should have the right to be informed, possibly through an interactive training program, on the areas of uncertainty, risks, and circumstances where safety or harm issues could arise, without this increasing user's accountability for A/IS decision-making consequences.

9. Lawmakers on national and 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 A/IS driven technology to prior testing and approval by appropriate national and/or international agencies. Companies should establish an A/IS ethics statement that includes statements about discrimination, addressing in that matter data-driven profiling and commitment to take measures to avoid user discrimination. In addition, companies should have internal systems that allow employees to identify and escalate issues related to discrimination in data and A/IS. Laws should create whistleblower protection for those who can and wish to reveal explicit violation of discrimination law. In particular, a well-crafted law to protect whistleblowers and to allow a public interest cause of action would improve accountability and aid in prevention of intentional misuse of A/IS.

10. The general public should be informed when articles/press releases related to political figures or issues are posted by an A/IS, such as a bot.

Further Resources

- Barocas, S., and A. D. Selbst. "Big Data's Disparate Impact." *California Law Review* 104 (2016): 671–732.
- Datta, A., S. Sen, and Y. 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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Law

- Etzioni, A., O. Etzioni. "Designing AI Systems That Obey Our Laws and Values." *Communications of the ACM* 59, no. 9 (2016): 29–31.
- Kroll, J. A., and J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. "Accountable Algorithms." (March 2, 2016). *University of Pennsylvania Law Review* 165 (2017 Forthcoming); *Fordham Law Legal Studies* Research Paper No. 2765268.
- Mittelstadt, B., P. Allo, M. Taddeo, S. Wachter, and L. Floridi. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* (July–December, 2016): 1–21.
- Regulation (EU) 2016/679 of the European Parliament and of the Council, General Data Protection Regulation (). "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." April 27, 2016.
- Wachter, S., B. Mittelstadt, and L. Floridi. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7, no. 2 (2017): 76–99.
- Zarsky, T. "The Trouble with Algorithmic Decisions: an Analytic Roadmap to Examine Efficiency and Fairness in Automated and Opaque Decision Making." *Science, Technology & Human Values* 41, no. 1 (2016): 118–132.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

Affect is a core aspect of intelligence. Drives and emotions, such as excitement and depression, are used to coordinate action throughout intelligent life, even in species that lack a nervous system. We are coming to realize that emotions are not an impediment to rationality, arguably they are integral to rationality in humans. Emotions are one evolved mechanism for satisficing — for getting what needs to be done in the time available with the information at hand. Emotions are core to how individuals and societies coordinate their actions. Humans are therefore susceptible to emotional influence both positively and negatively.

We would like to ensure that AI will be used to help humanity to the greatest extent possible in all contexts. In particular, artifacts used in society could cause harm either by amplifying or damping human emotional experience. It is quite possible we have reached a point where AI is affecting humans psychologically more than we realize. Further, even the rudimentary versions of synthetic emotions already in use have significant impact on how AI is perceived by policy makers and the general public.

This subcommittee addresses issues related to emotions and emotion-like control in both humans and artifacts. Our working groups have put forward candidate recommendations on a variety of concerns: considering how affect varies across human cultures, the particular problems of artifacts designed for intimate relations, considerations of how intelligent artifacts may be used for "nudging," how systems can support (or at least not interfere with) human flourishing, and appropriate policy concerning artifacts designed with their own affective systems.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

Document Sections

- Systems Across Cultures
- When Systems Become Intimate
- System Manipulation/Nudging/Deception
- Systems Supporting Human Potential (Flourishing)
- Systems with Their Own Emotions

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

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Systems Across Cultures

Issue:

Should affective systems interact using the norms appropriate for verbal and nonverbal communication consistent with the societal norms where they are located?

Background

Societies and therefore individuals around the world have different ways to maintain eye contact, express intentions through gestures, interpret silence, etc. These particularities could be incorporated into the affective systems in order to transmit the intended message. It would seem that an extensive study surrounding the norms/values of the community where the affective system will be deployed is essential to the system acceptability.

Candidate Recommendations

Any successful affective system should have a minimum set of ethical values/norms in its knowledge base that should be used in a specific cultural context. Some examples are listed below:

1. Affective systems should be careful in using small talk. Although small talk is useful for acting friendly, some communities see people that use small talk as insincere and hypocritical, while other cultures see the opposite and tend to consider people that do not use small talk as unfriendly, uncooperative, rude, arrogant, or ignorant. Additionally speaking with proper vocabulary, grammar, and sentence structure is often in contrast to the typical interactions that people have. In many mature economies, the latest trend, TV show, or other media can significantly influence what is viewed as appropriate vocabulary and interaction style.

2. Affective systems should recognize that the amount of personal space (proxemics) given is very important for human interaction. People from different cultures have different comfort zone distances to establish smooth communication. Crossing these limits without permission can transmit negative messages, such as hostile or sexual overtures.

3. Eye contact is an essential component in social interaction for certain cultures, while for others, it is not essential and may even generate misunderstandings or conflicts. It is important to recognize this in the development of such systems.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

4. Hand gestures and other non-verbal interaction are very important for social interaction, but should be used with caution across cultures and should be acknowledged in the design of affective systems. For instance, although a "thumbs-up" sign is commonly used to indicate approval, in some countries this gesture can be considered an insult.

5. Facial expressions are often used to detect emotions and facilitate emotional conversations. While it is tempting to develop A/IS that can recognize, analyze, and even display facial expressions for social interaction, it should be noted that facial expressions may not be universal across cultures and that an AI system trained with a dataset from one culture may not be readily usable in another culture.

Further Resources

The following documents/organizations can be used as additional resources to support the development of ethical affective systems.

- Cotton, G. "Gestures to Avoid in Cross-Cultural Business: In Other Words, 'Keep Your Fingers to Yourself!'" *Huffington Post,* June 13, 2013.
- "Cultural Intelligence & Paralanguage: Using Your Voice Differently Across Cultures." Sydney, Aus.: Culture Plus Consulting, 2016.
- Cotton, G. *Say Anything to Anyone, Anywhere: 5 Keys To Successful Cross-Cultural Communication*. Hoboken, NJ: Wiley, 2013.
- Elmer, D. *Cross-Cultural Connections: Stepping Out and Fitting In Around the World*. Westmont, IL: InterVarsity Press, 2002.
- Price, M. "Facial Expressions—Including Fear—May Not Be as Universal as We Thought." *Science,* October 17, 2016.

Issue:

Long-term interaction with affective artifacts lacking cultural sensitivity could alter the way people interact in society.

Background

Systems that do not have cultural knowledge incorporated into their knowledge base may change the way people interact, which may impact not only individuals, but also an entire society. Humans often use mirroring in order to understand and develop their principles and norms for behavior. At the same time, certain machine learning approaches focus on how to more appropriately interact with humans by mirroring human behavior. So learning via mirroring can go in both directions. If affective artifacts without cultural sensitivity interact with impressionable humans, they could alter the norms, principles, and therefore actions of that person. This creates a situation where the impact

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

of interacting with machines could significantly alter societal and cultural norms. For instance, children interacting with these systems can learn social and cultural values, which may be different from those present in their local community.

Candidate Recommendations

1. It is necessary to survey and analyze the long-term interaction of people with affective systems with different protocols and metrics to measure the modifications of habits, norms, and principles as well as the cultural and societal impacts.
2. Responsible parties (e.g., parents, nurse practitioners, social workers, and governments) should be trained to detect the influence due to AI and in effective mitigation techniques. In the most extreme case it should always be possible to shut down harmful A/IS.

Further Resources

The following documents can be used as guides to support the development of ethical affective systems.

- Nishida, T., and C. Faucher. *Modelling Machine Emotions for Realizing Intelligence: Foundations and Applications*. Berlin, Germany: Springer-Verlag, 2010.
- Pauleen, D. 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 (2006): 1–36.
- Bielby, J. "Comparative Philosophies in Intercultural Information Ethics." *Confluence: Online Journal of World Philosophies* 2, no. 1 (2015): 233–253.
- Bryson, J., "Why Robot Nannies Probably Won't Do Much Psychological Damage." A commentary on an article by N. Sharkey and A. Sharkey, *The Crying Shame of Robot Child Care Companions*. *Interaction Studies* 11, no. 2 (2010): 161–190.
- Sharkey, A., and N. Sharkey. "Children, the Elderly, and Interactive Robots." *IEEE Robotics & Automation Magazine* 18, no. 1 (2011): 32–38.

Issue:

When affective systems are inserted across cultures, they could affect negatively the cultural/socio/religious values of the community where they are inserted.

Background

Some philosophers believe there are no universal ethical principles; instead they argue that ethical norms vary from society to society. Regardless of whether universalism or some form of ethical relativism is true, affective systems need to respect the values of the cultures within which

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

they are embedded. To some it may be that we should be designing affective systems which can reflect the values of those with which the systems are interacting. There is a high likelihood that when spanning different groups, the values imbued by the developer will be different from the operator or customer of that affective system. Differences between affective systems and societal values can generate conflict situations (e.g., gestures being misunderstood, or prolonged or inadequate eye contact) that may produce undesirable results, perhaps even physical violence. Thus, affective systems should adapt to reflect the values of the community (and individuals) where they will operate in order to avoid conflict.

Candidate Recommendation

Assuming the affective systems have a minimum subset of configurable ethical values incorporated in their knowledge base:

1. They should have capabilities to identify differences between their values and the values of those they are interacting with and alter their interactions accordingly. As societal values change over time, any affective system needs to have the capability to detect this evolution and adapt its current ethical values to be in accordance with other people's values.
2. Those actions undertaken by an affective system that are most likely to generate an emotional response should be designed to be easily changed. Similar to how software today externalizes the language and vocabulary to be easily changed based on location, affective systems should externalize some of the core aspects of their actions.

Further Resources

The following documents/organizations can be used as guides to support the development of ethical affective systems.

- Bielby, J. "Comparative Philosophies in Intercultural Information Ethics." *Confluence: Online Journal of World Philosophies* 2, no. 1 (2015): 233–253.
- Velasquez, M., C. Andre, T. Shanks, and M. J. Meyer. "Ethical Relativism."Markkula Center for Applied Ethics, Santa Clara, CA: Santa Clara University, August 1, 1992.
- Culture reflects the moral values and ethical norms governing how people should behave and interact with others. "Ethics, an Overview." Boundless Management.
- Donaldson, T. "Values in Tension: Ethics Away from Home." *Harvard Business Review.* September–October 1996.
- The Center for Nonviolent Communication.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

When Systems Become Intimate

Issue:

Are moral and ethical boundaries crossed when the design of affective systems allows them to develop intimate relationships with their users?

Background

While robots capable of participating in an intimate relationship are not currently available, the idea that they could become intimate sexual partners with humans (e.g., sex robots) is one that captures the attention of the public and the media. Because the technology is already drawing much ethical scrutiny and may raise significant ethical concerns, it is important that policy makers and the professional community participate in developing guidelines for ethical research in this area. Part of the goal is to highlight potential ethical benefits and risks that may emerge, if and when affective systems develop intimacy with their users. Robots for use in the sex industry may help lessen human trafficking and the spread of STIs, but there is also the possibility that these systems could negatively impact human-to-human intimate relations. Human-to-human relations are currently viewed as being more rewarding, but also much more difficult to maintain than, for example, use of future robotic sex workers.

Candidate Recommendation

As this technology develops, it is important to monitor research in this realm and support those projects that enhance the user's development of intimate relationships in positive and therapeutic ways while critiquing those that contribute to problematic intimate relations, specifically:

1. Intimate systems must not be designed or deployed in ways that contribute to sexism, negative body image stereotypes, gender or racial inequality.
2. Intimate systems must avoid the sexual/psychological manipulation of the users of these systems unless the user is made aware they are being manipulated in this way (opt-in).
3. Intimate systems should not be designed in a way that contributes to user isolation from other human companions.
4. Designers of affective robotics, especially intimate systems, must foresee and publicly acknowledge that these systems can interfere with the relationship dynamics between human partners, causing jealousy or feelings of disgust to emerge between human partners.
5. Intimate systems must not foster deviant or criminal behavior. Sex robots should not be built in ways that lead to the normalization of taboo, unethical, or illegal sexual practices, such as pedophilia or rape.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

6. Commercially marketed AI should not be considered to be a person in a legal sense, nor marketed as a person. Rather its artifactual (authored, designed, and built deliberately) nature should always be made as transparent as possible, at least at point of sale and in available documentation (as noted in the Systems Supporting Human Potential Section below).

Further Resources

The following documents/organizations are provided for further research.

- Levy, D. *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: HarperCollins Publishers, 2007
- Scheutz, M. "The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots," in *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by P. Lin, K. Abney, and G. Bekey, 205. Cambridge, MA: MIT Press, 2011.
- Richardson, K. "The Asymmetrical 'Relationship': Parallels Between Prostitution and the Development of Sex Robots." *ACM SIGCAS Newsletter, SIGCAS Computers & Society* 45, no. 3 (2015): 290–293.
- Sullins, J. P. "Robots, Love, and Sex: The Ethics of Building a Love Machine." *IEEE Transactions on Affective Computing* 3, no. 4 (2012): 398–409.
- Yeoman, I., and M. Mars. "Article Robots, Men and Sex Tourism." *Futures* 44, no. 4 (2012): 365–371.
- Campaign against Sex Robots.
- Whitby, B. "Do You Want a Robot Lover? The Ethics of Caring Technologies," in *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by P. Lin, K. Abney, and G. A. Bekey, 233–248. Cambridge, MA: MIT Press, 2012.
- Danaher, J., and N. McArthur. *Robot Sex: Sexual and Ethical Implications*. Cambridge, MA: MIT Press, 2017.

Issue:

Can and should a ban or strict regulations be placed on the development of sex robots for private use or in the sex industry?

Background

The very idea of sex robots has sparked controversy even before many of these systems have become available. At this time, sex robots tend to be expensive love dolls made of silicone placed over a metal skeleton. These dolls can include robotic systems such as heating elements, sensors, movement, and rudimentary AI. The current state of the technology is a far cry from the sex robots imagined in novels and other media but they may just be the first step toward more advanced systems. There is ongoing debate around these systems. Critics are calling for strict

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

regulation or even a full ban on the development of this technology, while others argue that social value could be found by developing intimate robots, including on religious grounds.

Sex robots are already used for prostitution and this trend is likely to continue in many regions of the world. Some researchers report that robot prostitutes will completely revolutionize the sex tourism industry by 2050. For example, by that time, Amsterdam's Red Light District may be dominated by a variety of android systems with various capabilities (Yeoman and Mars, 2012). However there are critics of the technology, including those who are calling for an outright ban.

Despite being illegal, prostitution commonly occurs in many societies. Yet it is rarely done without creating deep ethical problems for the sex workers themselves and the societies in which they operate. Sex robots may alleviate some of these ethical concerns; for instance it has been argued that:

1. Sex robots might be less likely to be a vector for the transmission of sexually transmitted infections (STIs).
2. Sex robots could greatly lessen human trafficking of sex workers.
3. Sex robots could be regulated by policies on controlling prices, hours of operations, sexual services, and other aspects of prostitution.

However the technology can create serious ethical problems such as:

1. This technology would likely further normalize the sex industry, and that typically means a further tendency to objectify women, given that the majority of clients for these technologies are heterosexual men.
2. The availability of the technology could disrupt intimate relationships between human beings.

Human sexuality is an important human activity, but it comes associated with difficult ethical issues related to power and desire. This means that robot sexual partners will always be an ethically contentious technology. A comprehensive/global ban on sex robots is unlikely given that a large market for these technologies may already exist and is part of the current demand for sex toys and devices. However, there are important issues/considerations that the designers of these technologies need to consider.

Candidate Recommendation

1. We recommend regulation, not a ban, in accordance with cultural norms.
2. Existing laws regarding personal imagery need to be reconsidered in light of robot sexuality.
3. If it is proven through scientific studies that therapeutic uses of this technology could reduce recidivism in those who commit sex crimes, controlled use for those purposes only should be permitted, under legal and/or medical supervision.
4. Robot prostitution and sex tourism need to be monitored and controlled to fit local laws and policies.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

Further Resources

- Danaher, J., and N. McArthur. *Robot Sex: Sexual and Ethical Implications*. Cambridge, MA: MIT Press, 2017.
- Richardson, K. "The Asymmetrical 'Relationship': Parallels Between Prostitution and the Development of Sex Robots." *ACM SIGCAS Newsletter, SIGCAS Computers & Society* 45, no. 3 (2015): 290–293.
- Sullins, J. P. "Robots, Love, and Sex: The Ethics of Building a Love Machine." *IEEE Transactions on Affective Computing* 3, no. 4 (2012): 398–409.
- Yeoman, I., and M. Mars. "Robots, Men and Sex Tourism." *Futures* 44, no. 4 (2012): 365–371.
- Campaign against Sex Robots.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

System Manipulation/Nudging/Deception

Issue:

Should affective systems be designed to nudge people for the user's personal benefit and/or for the benefit of someone else?

Background

Emotional manipulation can be defined as an exercise of influence, with the intention to seize control and power at the person's expense. Thaler and Sunstein (2008) call the tactic of subtly modifying behavior a "nudge." Nudging mainly operates through the affective system. Making use of a nudge might be considered appropriate in situations like teaching children, treating drug dependency, healthcare, and when the global community benefits surpass individual benefits. Yet should affective systems be deployed to influence a user's behavior for that person's own good? Nudging can certainly trigger behaviors that worsen human health, but could the tactic be used by affective systems to cue behaviors that improve it? Several applications are possible in health, well-being, education, etc. Yet a nudge could have opposite consequences on different people, with different backgrounds and preferences (White, 2013, de Quintana Medina and Hermida Justo, 2016). Another key, and potentially more controversial, issue to be addressed is whether an affective system should be designed to nudge a user, and potentially intrude on individual liberty, when doing so may benefit someone else.

Candidate Recommendations

1. Systematic analyzes are needed that examine the ethics of designing affective systems to nudge human beings prior to deployment.
2. We recommend that the user be able to recognize and distinguish between different types of nudges, including ones that seek to promote beneficial social manipulation (e.g., healthy eating) versus others where the aim is psychological manipulation or the exploitation of an imbalance of power (e.g., for commercial purposes).
3. Since nudging alters behavior implicitly, the resulting data on infantilization effects should be collected and analyzed.
4. Nudging in autonomous agents and robots must have an opt-in system policy with explicit consent.
5. Additional protections must be put in place for vulnerable populations, such as children, when informed consent cannot be obtained, or when it may not be a sufficient safeguard.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

6. Nudging systems must be transparent and accountable, implying that data logging is required. This should include recording the user responses when feasible.

Further Resources

The following documents/organizations can be used as additional resources to support the development of ethical affective systems.

- Thaler, R., and C. R. Sunstein. *Nudge: Improving Decision about Health, Wealth and Happiness*, New Haven, CT: Yale University Press, 2008.
- Bovens, L. "The Ethics of Nudge," in *Preference change: Approaches from Philosophy, Economics and Psychology*, edited by T. Grüne-Yanoff and S. O. Hansson, 207–219. Berlin, Germany: Springer.
- de Quintana Medina, J., and P. Hermida Justo. "Not All Nudges Are Automatic: Freedom of Choice and Informative Nudges." Working paper presented to the European Consortium for Political Research, Joint Session of Workshops, 2016 Behavioral Change and Public Policy, Pisa, Italy, 2016.
- White, M. D. *The Manipulation of Choice. Ethics and Libertarian Paternalism.* New York: Palgrave Macmillan, 2013.
- Scheutz, M. "The Affect Dilemma for Artificial Agents: Should We Develop Affective Artificial Agents?" *IEEE Transactions on Affective Computing* 3, no. 4 (2012): 424–433.
- Grinbaum, A., R. Chatila, L. Devillers, J.-G. Ganascia, C. Tessier, and M. Dauchet. "Ethics in Robotics Research: CERNA Recommendations," *IEEE Robotics and Automation Magazine* 24, no. 3 (2017): 139–145.
- "Designing Moral Technologies: Theoretical, Practical, and Ethical Issues" Conference July 10–15, 2016, Monte Verità, Switzerland.

Issue:

Governmental entities often use nudging strategies, for example to promote the performance of charitable acts. But the practice of nudging for the benefit of society, including through the use of affective systems, raises a range of ethical concerns.

Background

A profoundly controversial practice that could be on the horizon is allowing a robot or another affective system to nudge a user for the good of society (Borenstein and Arkin, 2016). For instance, if it is possible that a well-designed robot could effectively encourage humans to perform charitable acts, would it be ethically appropriate for the robot to do so? This design possibility illustrates just one behavioral outcome that a robot could potentially elicit from a user.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

Given the persuasive power that an affective system may have over a user, ethical concerns related to nudging must be examined. This includes the significant potential for misuse.

Candidate Recommendations

As more and more computing devices subtly and overtly influence human behavior, it is important to draw attention to whether it is ethically appropriate to pursue this type of design pathway. There needs to be transparency regarding who the intended beneficiaries are, and whether any form of deception or manipulation is going to be used to accomplish the intended goal.

Further Resources

The following documents/organizations can be used as guides to support the development of ethical affective systems.

- Borenstein, J., and R. Arkin. "Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being." *Science and Engineering Ethics* 22, no. 1 (2016): 31–46.
- Borenstein, J., and R. C. Arkin. "Nudging for Good: Robots and the Ethical Appropriateness of Nurturing Empathy and Charitable Behavior." *AI and Society* 32, no. 4 (2016): 499–507.

Issue:

A nudging system that does not fully understand the context in which it is operating may lead to unintended consequences.

Background

This kind of system needs to have sophisticated enough technical capabilities for recognizing the context in which it is applying nudging strategies. We could imagine a technical license ("permits") (Omohundro, 2013).

Candidate Recommendation

1. When addressing whether affective systems should be permitted to nudge human beings, user autonomy is a key and essential consideration that must be taken into account.
2. We recommend that when appropriate, an affective system that nudges human beings should have the ability to accurately distinguish between users, including detecting characteristics such as whether the user is an adult or a child.
3. Affective systems with nudging strategies should be carefully evaluated, monitored, and controlled.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

Further Resources

The following documents/organizations can be used as guides to support the development of ethical affective systems.

- Borenstein, J., and R. Arkin. "Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being." *Science and Engineering Ethics* 22, no. 1 (2016): 31–46.
- Arkin, R. C., M. Fujita, T. Takagi, and R. Hasegawa. "An Ethological and Emotional Basis for Human-Robot Interaction." *Robotics and Autonomous Systems* 42, no. 3–4 (2003): 191–201.
- Omohundro, S. "Autonomous Technology and the Greater Human Good." *Journal of Experimental and Theoretical Artificial Intelligence* 26, no. 3 (2014): 303–315.

Issue:

When, if ever, and under which circumstances is deception performed by affective systems acceptable?

Background

Deception is commonplace in everyday human-human interaction. According to Kantian ethics, it is never ethically appropriate to lie, while utilitarian frameworks would indicate that it can be acceptable when it increases overall happiness. Given the diversity of views on the ethical appropriateness of deception, how should affective systems be designed to behave?

Candidate Recommendations

It is necessary to develop recommendations regarding the acceptability of deception in the design of affective autonomous agents with respect to when and under which circumstances, if any, it is appropriate.

1. In general, deception is acceptable in an affective agent when it is used for the benefit of the person being deceived, not for the agent itself. For example, deception might be necessary in search and rescue operations, elder- or child-care.
2. For deception to be used under any circumstance, a logical and reasonable justification must be provided by the designer, and this rationale must be approved by an external authority.
3. Deception must follow an opt-in strategy and must be transparent to the user, i.e., the context under which the system is allowed to deceive.

Further Resources

- Arkin, R. C., "Robots That Need to Mislead: Biologically-inspired Machine Deception." *IEEE Intelligent Systems* 27, no. 6 (2012): 60–75.
- Shim, J., and R. C. Arkin. "Other-Oriented Robot Deception: How Can a Robot's Deceptive Feedback Help Humans in HRI?"

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

Eighth International Conference on Social Robotics (ICSR 2016), Kansas, Mo., November 2016.

- Shim, J., and R. C. Arkin. "The Benefits of Robot Deception in Search and Rescue: Computational Approach for Deceptive Action Selection via Case-based Reasoning." *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR 2015)*, West Lafayette, IN, October 2015.
- Shim, J., and R. C. Arkin. "A Taxonomy of Robot Deception and its Benefits in HRI." *Proceedings of IEEE Systems, Man and Cybernetics Conference,* Manchester England, October 2013.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Systems Supporting Human Potential (Flourishing)

Issue:

Extensive use of artificial intelligence in society may make our organizations more brittle by reducing human autonomy within organizations, and by replacing creative, affective, empathetic components of management chains.

Background

As human workers are replaced by AI, their former employers (e.g., corporations and governments) may find they have eliminated the possibility of employees and customers discovering new equilibria outside the scope of what the organizations' leadership originally foresaw. Even in ordinary, everyday work, a lack of empathy based on shared needs and abilities disadvantages not only the liberty of individuals but also the corporations and governments that exist to serve them, by eliminating opportunities for useful innovation. Collaboration requires sufficient commonality of collaborating intelligences to create empathy — the capacity to model the other's goals based on one's own.

Candidate Recommendations

1. It is important that human workers within an organization have direct interactions with each other, rather than always being intermediated by affective systems (or other technology) which may filter out useful, unexpected communication. Similarly, we recommend human points of contact be available to customers and other organizations.

2. In particular, although there will be many cases where AI is less expensive, more predictable, and easier to control than human employees, we recommend maintaining a core number of human employees at every level of decision-making with clear communication pathways.

3. More generally, management and organizational theory should be extended to consider appropriate use of affective and autonomous systems to enhance their business model and the efficacy of their workforce.

Further Resource

The following document can be used as an additional resource to support the development of ethical affective systems.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

- Bryson, J. J. "Artificial Intelligence and Pro-Social Behavior," in *Collective Agency and Cooperation in Natural and Artificial Systems,* edited by Catrine Misselhorn, 281–306, Springer, 2015.

Issue:

The increased access to personal information about other members of our society, facilitated by artificial intelligence, may alter the human affective experience fundamentally, potentially leading to a severe and possibly rapid loss in individual autonomy.

Background

Theoretical biology tells us that we should expect increased communication — which AI facilitates — to increase group-level investment. This could have the effect of reducing individual autonomy and increasing in its place group-based identities. Candidate examples of this sort of social alteration include:

1. Increased investment in monitoring and controlling children's lives by parents.
2. Decreased willingness to express opinions for fear of surveillance or long-term unexpected consequences.
3. Utilization of "customers" to perform basic corporate business processes such as data entry as a barter for lower prices, resulting also in reduced tax revenues.

The loss of individual autonomy could lead to more fragmented or fragile societies, and (because diversity is associated with creativity) a reduction of innovation. This concern relates to issues of privacy and security, but also to social and legal liability for past expressions.

Candidate Recommendations

1. Organizations, including governments, must put a high value on individuals' privacy and autonomy, including restricting the amount and age of data held on individuals.
2. Educational countermeasures should be taken to encourage individuation and prevent loss of autonomy.

Further Resources

The following documents can be used as additional resources to support the development of ethical affective systems.

- Bryson, J. J. "Artificial Intelligence and Pro-Social Behavior," in *Collective Agency and Cooperation in Natural and Artificial Systems,* edited by Catrine Misselhorn, 281–306, Springer, 2015.
- Cooke, M.. "A Space of One's Own: Autonomy, Privacy, Liberty." *Philosophy & Social Criticism,* 25, no. 1, (1999): 22–53.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

- Roughgarden, J., M. Oishi, and E. Akçay. "Reproductive Social Behavior: Cooperative Games to Replace Sexual Selection." *Science* 311, no. 5763 (2006): 965–969.

Issue:

A/IS may negatively affect human psychological and emotional well-being in ways not otherwise foreseen.

Background

A/IS has unprecedented access to human culture and human spaces — both physical and intellectual — for something that is not a human. A/IS may communicate via natural language, it may move in humanlike forms, and express humanlike identity. As such, it may affect human well-being in ways not yet anticipated.

Candidate Recommendations

We recommend vigilance and research for identifying situations where A/IS are already affecting human well-being, both positively and negatively. We should look for evidence such as correlations between the increased use of A/IS and any suspected impacts. However, we should not be paranoid nor assume that correlation indicates causation. We recommend robust, ongoing, multidisciplinary research.

Further Resource

The following document can be used as an additional resource to support the development of ethical affective systems.

- Kamewari, K., M. Kato, T. Kanda, H. Ishiguro, and K. Hiraki. "Six-and-a-Half-Month-Old Children Positively Attribute Goals to Human Action and to Humanoid-Robot Motion." *Cognitive Development* 20, no. 2, (2005): 303–320.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Systems With Their Own Emotions

Issue:

Synthetic emotions may increase accessibility of AI, but may deceive humans into false identification with AI, leading to overinvestment of time, money, trust, and human emotion.

Background

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 A/IS, and therefore fail to realize that — unlike in evolved intelligence — synthetic emotions can be compartmentalized and even entirely removed. Potential consequences are over-bonding, guilt, and above all, misplaced trust. Because there is no coherent sense in which designed and engineered AI can be made to suffer, because any such affect, even if possible, could be avoided at the stage of engineering, or reengineered. Consequently, AI cannot be allocated moral agency or responsibility in the senses that have been developed for human sociality.

Candidate Recommendations

1. Commercially marketed AI should not be considered to be a person in a legal sense, nor marketed as a person. Rather its artifactual (authored, designed, and built deliberately) nature should always be made as transparent as possible, at least at point of sale and in available documentation.

2. Some systems will, due to their application, require opaqueness in some contexts (e.g., emotional therapy). Transparency in such instances should not be necessarily during operation, but the systems' working should still be available to inspection by responsible parties.

Further Resources

The following documents can be used as additional resources to support the development of ethical affective systems.

- Arkin, R. C., P. Ulam, and A. R. Wagner. "Moral Decision-making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust and Deception," *Proceedings of the IEEE* 100, no. 3 (2012): 571–589.

- Arkin, R., M. Fujita, T. Takagi, and R. Hasegawa. "An Ethological and Emotional Basis for Human-Robot Interaction," *Robotics and Autonomous Systems* 42, no. 3–4 (2003): 191–201.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Affective Computing

- Arkin, R. C.. "Moving Up the Food Chain: Motivation and Emotion in Behavior-based Robots," in *Who Needs Emotions: The Brain Meets the Robot*, edited by J. Fellous and M. Arbib. New York: Oxford University Press, 2005.
- Boden, M., J. Bryson, D. Caldwell, K. et al. "Principles of Robotics: Regulating Robots in the Real World." *Connection Science* 29, no. 2 (2017): 124–129.
- Bryson, J. J., M. E. Diamantis, and T. D. Grant. "Of, For, and By the People: The Legal Lacuna of Synthetic Persons." *Artificial Intelligence & Law* 25, no. 3 (2017): 273–291.
- Novikova, J., and L. Watts. "Towards Artificial Emotions to Assist Social Coordination in HRI." *International Journal of Social Robotics* 7 no. 1, (2015): 77–88.
- Scheutz, M. "The Affect Dilemma for Artificial Agents: Should We Develop Affective Artificial Agents?" *IEEE Transactions on Affective Computing* 3 (2012): 424–433.
- Sharkey, A., and N. Sharkey. "Children, the Elderly, and Interactive Robots." *IEEE Robotics & Automation Magazine* 18.1 (2011): 32–38.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Policy

Autonomous and intelligent systems (A/IS) are a part of our society. The use of these new, powerful technologies promotes a range of social goods, and may spur development across the economies and society through its numerous applications, including in commerce, employment, healthcare, transportation, politics, privacy, public safety, national security, civil liberties, and human rights. To protect the public from adverse consequences, intended or otherwise, resulting from these applications, effective A/IS public policies and government regulations are needed.

The goals of an effective A/IS policy center on the protection and promotion of safety, privacy, intellectual property rights, human rights, and cybersecurity, as well as the public understanding of the potential impact of A/IS on society. Without policies designed with these considerations in mind, there may be critical technology failures, loss of life, and high-profile social controversies. Such events could engender policies that unnecessarily stifle entire industries, or regulations that do not effectively advance public interest and protect human rights.

To ensure that A/IS best serves the public interest, we believe that effective A/IS policies should embody a rights-based approach[1] that achieves five principal objectives:

1. Support, promote, and enable internationally recognized legal norms
2. Develop workforce expertise in A/IS technology
3. Include ethics as a core competency in research and development leadership
4. Regulate A/IS to ensure public safety and responsibility
5. Educate the public on societal impacts of A/IS

1 This approach is rooted in internationally recognized economic, social, cultural, and political rights.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Policy

As autonomous and intelligent systems (A/IS) become a greater part of our everyday lives, managing the associated risks and rewards will become increasingly important. Technology leaders and policy makers have much to contribute to the debate on how to build trust, prevent drastic failures, and integrate ethical and legal considerations into the design of A/IS technologies.

Disclaimer: *While we have provided recommendations in this document, it should be understood these are not formal policy recommendations endorsed by IEEE and do not represent a position or the views of IEEE but the informed opinions of Policy Committee members providing insights designed to provide expert directional guidance regarding A/IS. In no event shall IEEE or IEEE-SA Industry Connections Activity Members be liable for any errors or omissions, direct or otherwise, however caused, arising in any way out of the use of this work, regardless of whether such damage was foreseeable.*

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Policy

Objective:

Ensure that A/IS support, promote, and enable internationally recognized legal norms.

Background

A/IS technologies have the potential to negatively impact internationally recognized economic, social, cultural, and political rights, through unintended outcomes or outright design decisions (as is the case with certain unmanned aircraft systems (Bowcott, 2013). In addition to the military application of A/IS, the domestic use of A/IS in predictive policing (Shapiro, 2017), banking (Garcia, 2017), judicial sentencing (Osoba and Welser, 2017), job hunting and hiring practices (Datta, Tschantz, and Datta, 2014), and even service delivery of goods (Ingold and Soper, 2016) can negatively impact human rights by automating certain forms of discrimination, inhibiting the right to assembly, freedom of expression, and access to information. To ensure A/IS are used as a force for good, it is crucial to formulate policies that prevent such violations of political, social, economic, and cultural rights.

A/IS regulation, development, and deployment should, therefore, be based on international human rights standards and standards of international humanitarian laws (in the case of armed conflicts). This can be achieved if both states and private actors consider their responsibility to respectively protect and respect internationally recognized political, social, economic, and cultural rights. For business actors, this means considering their obligation to respect international human rights, as laid out in the UN Guiding Principles for Business and Human Rights (OHCHR, 2011), also known as the Ruggie principles.

When discussing the responsibility of private actors, the UN Guiding Principles on Business and Human Rights should be reflected. These principles have been widely referenced and endorsed by corporations and led to the adoption of several corporate social responsibility (CSR) policies in various companies. As such, they have led to a better understanding of the role of businesses in protection and promotion of human rights and ensured that the most crucial human values and legal standards of human rights are respected by A/IS technologists.

Candidate Recommendations

A rights-based approach means using the internationally recognized legal framework for human rights standards that is directed at accounting for the impact of technology on individuals. This framework also addresses inequalities, discriminatory practices, and the unjust distribution of resources. A/IS right-based policies will reflect the following principles:

- Responsibility: The rights-based approach shall identify the right holders and the duty bearers, and ensure that duty bearers have an obligation to realize all human rights; this should guide the policy development and implementation of A/IS.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Policy

- Accountability: As duty bearers, states should be obliged to behave responsibly, seek to represent the greater public interest, and be open to public scrutiny of their A/IS policy.
- Participation: the rights-based approach demands a high degree of participation of all interested parties.
- Non-discrimination: Principles of non-discrimination, equality, and inclusiveness should underlie the practice of A/IS. The rights-based approach should also ensure that particular focus is given to vulnerable groups, to be determined locally, such as minorities, indigenous peoples, or persons with disabilities.
- Empowerment: The rights-based approach to A/IS should empower right holders to claim and exercise their rights.
- Corporate responsibility: Companies must ensure that when they are developing their technologies based on the values of a certain community, they do so only to the extent that such norms or values fully comply with the rights-based approach. Companies must also not willingly provide A/IS technologies to actors that will use them in ways that lead to human rights violations.

Further Resources

- Human rights-based approaches have been applied to development, education and reproductive health. See: the UN Practitioners' Portal on Human Rights Based Programming.
- Bowcott, O. "Drone Strikes By Us May Violate International Law, Says UN." *The Guardian*, October 18, 2013.
- Shapiro, A."Reform Predictive Policing." *Nature News* 541, no. 7638 (2017): 458.
- Garcia, M. "How to Keep Your AI from Turning Into a Racist Monster." *Wired*, April 21, 2017.
- Osoba, O. A., and W. Welser. "An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence." Santa Monica, CA: RAND Corporation, 2017.
- Datta, A., M. C. Tschantz, and A. Datta. "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination." arXiv:1408.6491 [Cs], 2014.
- Ingold, D., and S. Soper. "Amazon Doesn't Consider the Race of Its Customers. Should It?" *Bloomberg,* April 21, 2016.
- United Nations. *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*. United Nations Office of the High Commissioner of Human Rights. New York and Geneva: UN, 2011.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Policy

Objective:

Develop and make available to government, industry, and academia a workforce of well-qualified A/IS personnel.

Background

There is a clear consensus among private sector and academic stakeholders that effectively governing A/IS and related technologies requires a level of technical expertise that governments currently do not possess. Effective governance requires more experts who understand and can analyze the interactions between A/IS technologies, programmatic objectives, and overall societal values. With current levels of technical understanding and expertise, policies and regulations may fail to support innovation, adhere to national principles, and protect public safety.

At the same time, the A/IS personnel should not only possess a necessary technology knowledge, but also receive adequate ethical training, and have access to other resources on human rights standards and obligations, along with guidance on how to make them a fundamental component of their work.

Candidate Recommendations

A high level of technical expertise is required to create a public policy, legal, and regulatory environment that allows innovation to flourish while protecting the public and gaining public trust.[1] Policy makers and market leaders should pursue several strategies for developing this expertise:

- Expertise can be furthered by setting up technical fellowships, or rotation schemes, where technologists spend an extended time in political offices, or policy makers work with organizations that operate at the intersection of tech-policy, technical engineering, and advocacy (like the American Civil Liberties Union, Article 19, the Center for Democracy and Technology, or Privacy International). This will enhance the technical knowledge of policy makers and strengthen ties between political and technical communities, needed to make good A/IS policy.
- A culture of sharing best practices around A/IS legislation, consumer protection, workforce transformation, and economic displacement stemming from A/IS-based automation should be fostered across borders. This can be done by doing exchange governmental delegation trips, transcontinental knowledge exchanges, and by building A/IS components into existing venues and efforts surrounding good regulation (General Data Protection Regulation (GDPR)).

1 This recommendation concurs with the multiple recommendations of the United States National Science and Technology Council, One Hundred Year Study of Artificial Intelligence, Japan's Cabinet Office Council, European Parliament's Committee on Legal Affairs and others.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Policy

- In order to ensure that the next generation of policy makers is tech savvy, it is necessary to rely upon more than their "digital nativeness." Because A/IS are evolving technologies, long-term educational strategies are needed, e.g., providing children access to coding and computer science courses starting from primary school, and extending into university or vocational courses.

Further Resources

- Holdren, J., and M. Smith. "Preparing for the Future of Artificial Intelligence." Washington, DC: Executive Office of the President, National Science and Technology Council, 2016.
- Stanford University. "Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence." Stanford, CA: Stanford University, 2016.
- "Japan Industrial Policy Spotlights AI, Foreign Labor." *Nikkei Asian Review,* May 20, 2016.
- Weng, Y.-H. "A European Perspective on Robot Law: Interview with Mady Delvaux-Stehres." *Robohub*, July 15, 2016.

Objective:

Support research and development needed to ensure continued leadership in A/IS.

Background

Greater national investment in ethical A/IS research and development would stimulate the economy, create high-value jobs, and improve governmental services to society. A/IS can significantly improve our societies: the use of A/IS in computer vision and human-computer interactions will have far-reaching implications. Intelligent robots will perform difficult and dangerous tasks that require human-like intelligence. Self-driving cars will revolutionize automobile transportation and logistics systems and reduce traffic fatalities. A/IS will improve quality of life through smart cities and decision support in healthcare, social services, criminal justice, and the environment. However, to ensure such a positive impact, more support for R&D, with a particular eye for the ethical impact of A/IS, is needed.

Candidate Recommendations

Investment in A/IS research and development (including ethical considerations) is essential to maximizing societal benefits, mitigating any associated risks, and enabling efficient and effective public sector investment. To enable efficient and effective public and private sector investment, there should be benchmarks

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Policy

for A/IS technologies and applications with continuing focus on identifying promising future applications of A/IS. An important government role is to strategically educate the public and private sectors on key A/IS technologies and applications. We recommend the following:

- Enable a cross-disciplinary research environment that encourages research on the fairness, security, transparency, understandability, privacy, and societal impacts of A/IS and that incorporates independent means to properly vet, audit, and assign accountability to the A/IS applications.
- Governments should create research pools that incentivize research on A/IS that benefits the public, but which may not be commercially viable.

Further Resources

- Kim, E. T. "How an Old Hacking Law Hampers the Fight Against Online Discrimination." *The New Yorker*, October 1, 2016.
- National Research Council. "*Developments in Artificial Intelligence, Funding a Revolution: Government Support for Computing Research.*" Washington, DC: National Academy Press, 1999.
- Chen, N., L. Christensen, K. Gallagher, R. Mate, and G. Rafert (Analysis Group). "Global Economic Impacts of Artificial Intelligence," February 25, 2016.
- The Networking and Information Technology Research and Development Program, "Supplement to the President's Budget, FY2017." NITRD National Coordination Office, April 2016.
- Furber, S. B., F. Galluppi, S. Temple, and L. A. Plana. "The SpiNNaker Project." *Proceedings of the IEEE* 102, no. 5 (2014): 652–665.
- Markram, H. "The Human Brain Project." *Scientific American* 306, no. 2 (June 2012): 50–55.
- L. Yuan. "China Gears Up in Artificial-Intelligence Race." *Wall Street Journal*, August 24, 2016.

Objective:

Provide effective regulation of A/IS to ensure public safety and responsibility while fostering a robust AI industry.

Background

Governments must ensure consistent and appropriate policies and regulations for A/IS. Effective regulation should address transparency, understandability, predictability, and accountability of AI algorithms, risk management, data protection, and safety. Certification of systems involving A/IS is

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Policy

a key technical, societal, and industrial issue. Good regulation encourages innovation, and harmonizing policy internationally will reduce barriers to trade.

Good regulation can take many different forms, and appropriate regulatory responses are context-dependent. There is no one-size-fits-all for A/IS regulation, but it is important that such regulation is developed through an approach that is based on human rights[2] and has human well-being as a key goal.

Candidate Recommendations

- To ensure consistent and appropriate policies and regulations across governments, policymakers should seek informed input from a range of expert stakeholders, including academic, industry, and government officials, to consider questions related to the governance and safe employment of A/IS.
- To foster a safe international community of A/IS users, policymakers should take similar work being carried out around the world into consideration. Due to the transnational nature of A/IS, globally synchronized policies can have a greater impact on public safety and technological innovation.
- Law schools should offer interdisciplinary courses such as "Introduction to AI and Law" to reduce the gap between regulators, lawyers, and A/IS researchers and developers.
- Establish policies that foster the development of economies able to absorb A/IS, while providing broad job opportunities to those who might otherwise be alienated or unemployed. In addition, the continued development of A/IS talent should be fostered through international collaboration.
- Continue research into the viability of universal basic income. Such a non-conditional and government-provided addition to people's income might lighten the economic burden that comes from automation and economic displacement caused by A/IS.
- Ambiguity regarding whether and how proprietary A/IS may be reverse engineered and evaluated by academics, journalists, and other researchers can stifle innovation and public safety. Elimination of these impediments is essential.

Further Resources

- Stanford University. "Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence." Stanford, CA: Stanford University, 2016.
- Calo, R. "The Case for a Federal Robotics Commission." The Brookings Institution, 2014.
- Mannes, A. "Institutional Options for Robot Governance," 1–40, in *We Robot 2016*, Miami, FL, April 1–2, 2016.

2 Human rights–based approaches have been applied to development, education, and reproductive health. See: the UN Practitioner's Portal on Human Rights Based Programming.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Policy

- Marchant, G. E., K. W. Abbott, and B. Allenby, *Innovative Governance Models for Emerging Technologies*. Cheltenham, U.K.: Edward Elgar Publishing, 2014.
- Weng, Y. H., Y. Sugahara, K. Hashimoto, and A. 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 (2015): 841–857.

Objective:

Facilitate public understanding of the rewards and risks of A/IS.

Background

Perception drives public response. A/IS technologies and applications can both capture the imagination such as self-driving cars, and instill fear. Therefore, it is imperative for industry, academia, and government to communicate accurately both the positive potential of A/IS and the areas that require caution. Developing strategies for informing and engaging the public on A/IS benefits and challenges are critical to creating an environment conducive to effective decision-making.

The success of A/IS technology depends on the ease with which people use and adapt to A/IS applications. While improving public understanding of A/IS technologies through education is becoming increasingly important, so is the need to educate the public about the social and cultural issues of A/IS. The way A/IS interact with final users, build cognitive models of their power and limits, and so help their adoption and sense of control, are key technological objectives.

If society approaches these technologies primarily with fear and suspicion, societal resistance may result, impeding important work on ensuring the safety and reliability of A/IS technologies. On the other hand, if society is informed of the positive contributions and the opportunities A/IS create, then the technologies emerging from the field could profoundly transform society for the better in the coming decades.[3]

Another major societal issue — and the subject of much ongoing debate — is whether A/IS should have, or could develop, any sense of ethical behavior. A/IS will require a commonly accepted sense of ethical behavior, or, at the very least, possess behaviors with ethical implications. Therefore, technology awareness *and* understanding of social and ethical issues of A/IS are new literacy skills society must embrace if A/IS applications are to be accepted and trusted as an integral part of modern living.

3 One hundred year study of AI (AI100), Stanford University, August, 2016.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Policy

Candidate Recommendations

- Encourage A/IS development to serve the pressing needs of humanity by promoting dialogue and continued debate over the social and ethical implications of A/IS. To better understand the societal implications of A/IS, we recommend that funding be increased for interdisciplinary research on topics ranging from basic research into intelligence to principles on ethics, safety, privacy, fairness, liability, and trustworthiness of A/IS technology. Societal aspects should be addressed not only at an academic level but also through the engagement of business, public authorities, and policy makers. While technical innovation is a goal, it should not be prioritized over the protection of individuals.

- Begin an international multi-stakeholder dialogue to determine the best practices for using and developing A/IS, and codify this dialogue into international norms and standards. Many industries, in particular system industries (automotive, air and space, defense, energy, medical systems, manufacturing) are going to be significantly changed by the surge of A/IS. A/IS algorithms and applications must be considered as products owned by companies, and therefore the companies must be responsible for the A/IS products not being a threat to humanity.

- Empower and enable independent journalists and media outlets to report on A/IS, both by providing access to technical expertise and funding for independent journalism.

- Conduct media outreach to illustrate A/IS beneficial uses, and the important steps being taken to ensure safety and transparency. Public opinion related to trust, safety, privacy, employment, and the economy will drive public policy. It is critical to creating an environment conducive to effective decision-making, particularly as more government services come to rely on A/IS, that strategies are developed to inform and engage the public on AI benefits and challenges. Care must be taken to augment human interaction with A/IS and to avoid discrimination against segments of society.

Further Resources

- Networking and Information Technology Research and Development (NITRD) Program. "The National Artificial Intelligence Research and Development Strategic Plan." Washington, DC: Office of Science and Technology Policy, 2016.

- Saunders, J., P. Hunt, and J. S. Hollywood. "Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot," *Journal of Experimental Criminology* 12, no. 347, (2016): 347–371. doi:10.1007/s11292-019272-0

- Edelman, B., and M. Luca. "Digital Discrimination: The Case of Airbnb.com." Harvard Business School Working Paper 14-054, 2014.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Policy

- Garvie, C., A. Bedoya, and J. Frankle. "The Perpetual Line-Up: Unregulated Police Face Recognition in America." Washington, DC: Georgetown Law, Center on Privacy & Technology, 2016.
- Chui M., and J. Manyika, "Automation, Jobs, and the Future of Work." Seattle, WA: McKinsey Global Institute, 2014.
- The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Artificial Intelligence and Autonomous Systems,* Version 1. IEEE, 2016.
- Arkin, R. C. "Ethics and Autonomous Systems: Perils and Promises [Point of View]." *Proceedings of the IEEE* 104, no. 10, (1779–1781): 2016.
- Eurobarometer Survey on Autonomous Systems (published June 2015 by DG Connect) looks at Europeans' attitudes to robots, driverless vehicles, and autonomous drones. The survey shows that those who have more experience with robots (at home, at work or elsewhere) are more positive toward their use.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

The task of the Committee for Classical Ethics in Autonomous and Intelligent Systems is to apply classical ethics methodologies to considerations of algorithmic design in autonomous and intelligent systems (A/IS) where machine learning may or may not reflect ethical outcomes that mimic human decision-making. To meet this goal, the Committee has drawn from classical ethics theories as well as from the disciplines of machine ethics, information ethics, and technology ethics.

As direct human control over tools becomes, on one hand, further removed, but on the other hand, more influential than ever through the precise and deliberate design of algorithms in self-sustained digital systems, creators of autonomous systems must ask themselves how cultural and ethical presumptions bias artificially intelligent creations, and how these created systems will respond based on such design.

By drawing from over two thousand years' worth of classical ethics traditions, the Classical Ethics in Autonomous and Intelligent Systems Committee will explore established ethics systems, addressing both scientific and religious approaches, including secular philosophical traditions such as utilitarianism, virtue ethics, and deontological ethics and religious-and-culture-based ethical systems arising from Buddhism, Confucianism, African Ubuntu traditions, and Japanese Shinto influences toward an address of human morality in the digital age. In doing so the Committee will critique assumptions around concepts such as good and evil, right and wrong, virtue and vice and attempt to carry these inquiries into artificial systems decision-making processes.

Through reviewing the philosophical foundations that define autonomy and ontology, the Committee will address the potential for autonomous capacity of artificially intelligent systems, posing questions of morality in amoral systems, and asking whether decisions made by amoral systems can have moral consequences. Ultimately, it will address notions of responsibility and accountability for the decisions made by autonomous systems and other artificially intelligent technologies.

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 1 — Definitions for Classical Ethics in Autonomous and Intelligent Systems Research

Issue:

Assigning foundations for morality, autonomy, and intelligence.

Background

Classical theories of economy in the Western tradition, starting with Plato and Aristotle, embrace three domains: the individual, the family, and the *polis*. The forming of the individual character (*ethos*) is intrinsically related to others, as well as to the tasks of administration of work within the family (*oikos*) and eventually all this expands into the framework of the *polis*, or public space (*poleis*). This means that when we discuss ethical issues of autonomous and intelligent systems we should consider all three traditional economic dimensions that evolved in modernity into an individual morality disconnected from economics and politics. This disconnection was partly questioned by thinkers such as Adam Smith, Hegel, Marx, and others. In particular, Immanuel Kant's ethics located morality within the subject (see: categorical imperative) and separated morality from the outside world and the consequences of being a part of the outside world. The moral autonomous subject of modernity became thus a worldless isolated subject. This process is important to understand in terms of ethics for artificial intelligence since it is, paradoxically, the kind of autonomy that is supposed to be achieved by intelligent machines in the very moment in which we, humans, begin to change our being into digitally networked beings.

There lies a danger in uncritically attributing classical concepts of anthropomorphic autonomy to machines, including using the term *artificial intelligence* to describe them since, in the attempt to make them "moral" by programming moral rules into their behavior, we run the risk of assuming economic and political dimensions that do not exist, or that are not in line with contemporary human societies. As noted above, present human societies are being redefined in terms of digital citizenship via digital social networks. The present public debate about the replaceability of human work by *intelligent* machines is a symptom of this lack of awareness of the economic and political dimensions as defined by classical ethics, reducing ethical thinking to the "morality" of a worldless and isolated machine (a mimic of the modern subject).

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

Candidate Recommendations

- Via a return to classical ethics foundations, enlarge the discussion on ethics in autonomous and intelligent systems (A/IS) to include a critical assessment of anthropomorphic presumptions of ethics and moral rules for A/IS. Keep in mind that machines do not, in terms of classical autonomy, comprehend the moral or legal rules they follow, but rather move according to what they are programmed to do, following rules that are designed by humans to be moral.
- Enlarge the discussion on ethics for A/IS to include an exploration of the classical foundations of economy, outlined above, as potentially influencing current views and assumptions around machines achieving isolated autonomy.

Further Resources

- Bielby, J., ed. "Digital Global Citizenship." *International Review of Information Ethics* 23 (November 2015).
- Bendel, O. "Towards a Machine Ethics." Northwestern Switzerland: University of Applied Sciences and Arts, 2013.
- Bendel, O. "Considerations about the Relationship Between Animal and Machine Ethics." *AI & Society* 31, no. 1 (2016): 103–108.
- Capurro, R., M. Eldred, and D. Nagel. Digital Whoness: Identity, Privacy and Freedom in the Cyberworld. Berlin: Walter de Gruyter, 2013.
- Chalmers, D. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17, (2010): 7–65.

Issue:

Distinguishing between agents and patients.

Background

Of concern for understanding the relationship between human beings and A/IS is the uncritically applied anthropomorphistic approach toward A/IS that many industry and policy makers are using today. This approach erroneously blurs the distinction between moral agents and moral patients (i.e., subjects), otherwise understood as a distinction between "natural" self-organizing systems and artificial, non-self-organizing devices. As noted above, A/IS devices cannot, by definition, become autonomous in the sense that humans or living beings are autonomous. With that said, autonomy in machines, when critically defined, designates how machines act and operate independently in certain contexts through a consideration of implemented order generated by laws and rules. In this sense, A/IS can, by definition, qualify as autonomous, especially in the case of genetic algorithms and evolutionary strategies. However, attempts

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

to implant true morality and emotions, and thus accountability (i.e., autonomy) into A/IS is both dangerous and misleading in that it encourages anthropomorphistic expectations of machines by human beings when designing and interacting with A/IS.

Thus, an adequate assessment of expectations and language used to describe the human-A/IS relationship becomes critical in the early stages of its development, where unpacking subtleties is necessary. Definitions of autonomy need to be clearly drawn, both in terms of A/IS and human autonomy. On one hand A/IS may in some cases manifest seemingly ethical and moral decisions, resulting for all intents and purposes in efficient and agreeable moral outcomes. Many human traditions, on the other hand, can and have manifested as fundamentalism under the guise of morality. Such is the case with many religious moral foundations, where established cultural mores are neither questioned nor assessed. In such scenarios, one must consider whether there is any functional difference between the level of autonomy in A/IS and that of assumed agency (the ability to choose and act) in humans via the blind adherence to religious, traditional, or habitual mores. The relationship between assumed moral customs (mores), the ethical critique of those customs (i.e., ethics), and the law are important distinctions.

The above misunderstanding in definitions of autonomy arise in part because of the tendency for humans to shape artificial creations in their own image, and our desire to lend our human experience to shaping a morphology of artificially intelligent systems. This is not to say that such terminology cannot be used metaphorically, but the difference must be maintained, especially as A/IS begins to resemble human beings more closely. Terms like "artificial intelligence" or "morality of machines" can be used as metaphors, and it does not necessarily lend to misunderstanding to do so. This is how language works and how humans try to understand their natural and artificial environment.

However the critical difference between human autonomy and autonomous systems involves questions of free will, predetermination, and being (ontology). The questions of critical ontology currently being applied to machines are not new questions to ethical discourse and philosophy and have been thoroughly applied to the nature of human *being* as well. John Stuart Mill, for example, is a determinist and claims that human actions are predicated on predetermined laws. He does, however, argue for a reconciliation of human free will with determinism through a theory of compatibility. Millian ethics provides a detailed and informed foundation for defining autonomy that could serve to help combat general assumptions of anthropomorphism in A/IS and thereby address the uncertainty therein (Mill, 1999).

Candidate Recommendation

When addressing the nature of "autonomy" in autonomous systems, it is recommended that the discussion first consider free will, civil liberty, and society from a Millian perspective in order to better grasp definitions of autonomy and to combat general assumptions of anthropomorphism in A/IS.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

Further Resources

- Capurro, Rafael. "Toward a Comparative Theory of Agents." *AI & Society* 27, no. 4 (2012): 479–488.
- King, William Joseph, and Jun Ohya. "The representation of agents: Anthropomorphism, agency, and intelligence." Conference Companion on Human Factors in Computing Systems. ACM, 1996.
- Hofkirchner, W. "Does Computing Embrace Self-Organization?" in *Information and Computation, Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation,* edited by G. Dodig-Crnkovic, M. Burgin, 185–202. London: World Scientific, 2011.
- International Center for Information Ethics.
- Mill, J. S. *On Liberty.* London: Longman, Roberts & Green, 1869.
- Verbeek, P.-P. What Things Do: Philosophical Reflections on Technology, Agency, and Design. University Park, PA: Penn State Press, 2010.

Issue:

There is a need for an accessible classical ethics vocabulary.

Background

Philosophers and ethicists are trained in vocabulary relating to philosophical concepts and terminology. There is an intrinsic value placed on these concepts when discussing ethics and AI, since the layered meaning behind the terminology used is foundational to these discussions, and is grounded in a subsequent entrenchment of values. Unfortunately, using philosophical terminology in cross-discipline instances, for example, in conversation with technologists and policymakers is often ineffective since not everyone has the education to be able to encompass the abstracted layers of meaning contained in philosophical terminology.

However, not understanding a philosophical definition does not detract from the necessity of its utility. While ethical and philosophical theories should not be over-simplified for popular consumption, being able to adequately translate the essence of the rich history of ethics traditions will go a long way in supporting a constructive dialogue on ethics and A/IS. As access and accessibility concerns are also intricately linked with education in communities, as well as secondary and tertiary institutions, society needs to take a vested interest in creating awareness

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

for government officials, rural communities, and school teachers. Creating a more "user-friendly" vocabulary raises awareness on the necessity and application of classical ethics to digital societies.

Candidate Recommendation

Support and encourage the efforts of groups raising awareness for social and ethics committees whose roles are to support ethics dialogue within their organizations, seeking approaches that are both aspirational and values-based. A/IS technologists should engage in cross-discipline exchanges whereby philosophy scholars and ethicists attend and present at non-philosophical courses. This will both raise awareness and sensitize non-philosophical scholars and practitioners to the vocabulary.

Further Resources

- Capurro, R. "Towards an Ontological Foundation of Information Ethics." *Ethics and Information Technology* 8, no. 4 (2006): 175–186.
- Flinders, D. J. "In Search of Ethical Guidance: Constructing a Basis for Dialogue 1." *Qualitative Studies in Education* 5, no. 2 (1992): 101–115.
- Saldanha, G. S. "The Demon in the Gap of Language: Capurro, Ethics and Language in Divided Germany." *Information Cultures in the Digital Age*. Wiesbaden, Germany: Springer Fachmedien, 2016. 253–268.

Issue:

Presenting ethics to the creators of autonomous and intelligent systems.

Background

The question arises as to whether or not classical ethics theories can be used to produce meta-level orientations to data collection and data use in decision-making. The key is to embed ethics into engineering in a way that does not make ethics a servant, but instead a partner in the process. In addition to an ethics-in-practice approach, providing students and engineers with the tools necessary to build a similar orientation into their devices further entrenches ethical design practices. In the abstract this is not so difficult to describe, but very difficult to encode into systems.

This problem can be addressed by providing students with job-aids such as checklists, flowcharts, and matrices that help them select and use a principal ethical framework, and then exercise use of those devices with steadily more complex examples. In such an iterative process, students will start to determine for themselves what examples do not allow for perfectly clear decisions, and in fact require some interaction between frameworks. Produced outcomes such as videos, essays, and other formats – such as project-based learning activities – allow for a didactical strategy which proves effective in artificial intelligence ethics education.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

The goal is to provide students a means to use ethics in a manner analogous to how they are being taught to use engineering principles and tools. In other words, the goal is to help engineers tell the story of what they're doing.

- Ethicists should use information flows and consider at a meta-level what information flows do and what they are supposed to do.
- Engineers should then build a narrative that outlines the iterative process of ethical considerations in their design. Intentions are part of the narrative and provide a base to reflect back on those intentions.
- The process then allows engineers to better understand their assumptions and adjust their intentions and design processes accordingly. They can only get to these by asking targeted questions.

This process, one with which engineers are quite familiar, is basically Kantian and Millian ethics in play.

The aim is to produce what in computer programming lexicon is referred to as a *macro*. A macro is code that takes other code as its input(s) and produces unique outputs. This macro is built using the Western ethics tradition of virtue ethics.

Candidate Recommendation

Find ways to present ethics where the methodologies used are familiar to engineering students. As engineering is taught as a collection of *techno-science, logic, and mathematics*, embedding ethical sensitivity into these objective and non-objective processes is essential.

Further Resources

- Bynum, T. W., and S. Rogerson. *Computer Ethics and Professional Responsibility.* Malden, MA: Wiley-Blackwell, 2003.
- Seebauer, E. G., and R. L. Barry. *Fundamentals of Ethics for Scientists and Engineers.* New York: Oxford University Press, 2001.
- Whitbeck, C. "Teaching Ethics to Scientists and Engineers: Moral Agents and Moral Problems." *Science and Engineering Ethics* 1, no. 3 (1995): 299–308.
- Zevenbergen, B. et al. "Philosophy Meets Internet Engineering: Ethics in Networked Systems Research." GTC workshop outcomes paper. Oxford, U.K.: Oxford Internet Institute, University of Oxford, 2015.
- Perez Á., and M. Ángel, "Teaching Information Ethics." *International Review of Information Ethics* 14 (12/2010): 23–28.
- Verbeek, P-P. *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press, 2011.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

Issue:

Access to classical ethics by corporations and companies.

Background

Many companies, from start-ups to tech giants, understand that ethical considerations in tech design are increasingly important, but are not quite sure how to incorporate ethics into their tech design agenda. How can ethical considerations in tech design become an integrated part of the agenda of companies, public projects, and research consortia? Many corporate workshops and exercises that attempt to consider ethics in technology practices present the conversation as a carte blanche for people to speak about their opinions, but serious ethical discussions are often lacking. As it stands, classical ethics is not accessible enough to corporate endeavors in ethics, and as such, are not applicable to tech projects. There is often, but not always, a big discrepancy between the output of engineers, lawyers, and philosophers when dealing with computer science issues and a large difference in how various disciplines approach these issues. While this is not true in all cases, and there are now several interdisciplinary approaches in robotics and machine ethics as well as a growing number of scientists that hold double and interdisciplinary degrees, there remains a vacuum for the wider understanding of classical ethics theories in the interdisciplinary setting.

Candidate Recommendation

Bridge the language gap between technologists, philosophers, and policymakers. Understanding the nuances in philosophical language is critical to digital society from IoT, privacy, and cybersecurity to issues of Internet governance.

Further Resources

- Bhimani, A. "Making Corporate Governance Count: The Fusion of Ethics and Economic Rationality." *Journal of Management & Governance* 12, no. 2 (2008): 135–147.
- Carroll, A. B. "A History of Corporate Social Responsibility." in *The Oxford Handbook of Corporate Social Responsibility,* edited by Chrisanthi A., R. Mansell, D. Quah, and R. Silverstone. Oxford, U.K.: Oxford University Press, 2008.
- Lazonick, W. "Globalization of the ICT Labor Force." in *The Oxford Handbook of Information and Communication Technologies,* edited by Chrisanthi A., R. Mansell, D. Quah, and R. Silverstone. Oxford, U.K.: Oxford University Press, 2006.
- IEEE P7000™, Model Process for Addressing Ethical Concerns During System Design. This standard will provide engineers and technologists with an implementable process aligning innovation management processes, IS system design approaches and software engineering methods to minimize ethical risk for their organizations, stakeholders and end users. The Working Group is currently in process, and is free and open to join.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

Issue:

Impact of automated systems on the workplace.

Background

The impact of A/IS on the workplace and the changing power relationships between workers and employers requires ethical guidance. Issues of data protection and privacy via big data in combination with the use of autonomous systems by employers is an increasing issue, where decisions made via aggregate algorithms directly impact employment prospects. The uncritical use of A/IS in the workplace in employee/employer relations is of utmost concern due to the high chance for error and biased outcome.

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

Quoting Von Schomberg, "Responsible Research and Innovation is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)."[1]

When RRI methodologies are used in the ethical considerations of A/IS design, especially in response to the potential bias of A/IS in the workplace, theoretical deficiencies are then often exposed that would not otherwise have been exposed, allowing room for improvement in design at the development stage rather than from a retroactive perspective. RRI in design increases the chances of both relevance and strength in ethically aligned design.

Candidate Recommendation

It is recommended that through the application of RRI, as founded in classical ethics theory, research in A/IS design utilize available tools and approaches to better understand the design process, addressing ethical concerns from the very beginning of the design stage of the project, thus maintaining a stronger more efficient methodological accountability throughout.

Further Resources

- Burget, M., E. Bardone, and M. Pedaste. "Definitions and Conceptual Dimensions of Responsible Research and Innovation: A Literature Review." *Science and Engineering Ethics* 23, no. 1 (2016): 1–9.

1 Von Schomberg (2011) 'Prospects for Technology Assessment in a framework of responsible research and innovation' in: M. Dusseldorp and R. Beecroft (eds). Technikfolgen abschätzen lehren: Bildungspotenziale transdisziplinärer Methoden, Wiesbaden: Vs Verlag, in print, P.9.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

- Von Schomberg, R. "Prospects for Technology Assessment in a Framework of Responsible Research and Innovation," in *Technikfolgen Abschätzen Lehren: Bildungspotenziale Transdisziplinärer Methode*, 39–61, Wiesbaden, Germany: Springer VS, 2011.
- Stahl, B. C. et al. "From Computer Ethics to Responsible Research and Innovation in ICT: The Transition of Reference Discourses Informing Ethics-Related Research in Information Systems." *Information & Management* 51, no. 6 (2014): 810–818.
- Stahl, B. C., and B. Niehaves. "Responsible Research and Innovation (RRI)."
- IEEE P7005™, Standard for Transparent Employer Data Governance is designed to provide organizations with a set of clear guidelines and certifications guaranteeing they are storing, protecting, and utilizing employee data in an ethical and transparent way. The Working Group is currently in process, and is free and open to join.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 2 — Classical Ethics From Globally Diverse Traditions

Issue:

The monopoly on ethics by Western ethical traditions.

Background

As human creators, our most fundamental values are imposed on the systems we design. It becomes incumbent on a global-wide community to recognize which sets of values guide the design, and whether or not A/IS will generate problematic (e.g., discriminatory) consequences without consideration of non-Western values. There is an urgent need to broaden traditional ethics in its contemporary form of "responsible innovation" (RI) beyond the scope of "Western" ethical foundations, e.g., utilitarianism, deontology, and virtue ethics; and include other traditions of ethics in RI, including those inherent to, for example, Buddhism, Confucianism, and Ubuntu traditions.

However, this venture poses problematic assumptions even before the issue above can be explored, when, in classifying Western values, we also group together thousands of years of independent and disparate ideas originating from the Greco-Roman philosophical tradition with its Christian-infused cultural heritage.

What is it that one refers to by the term *Western ethics*? By Western ethics, does one refer to philosophical ethics (ethics as a scientific discipline) or is the reference to Western morality?

The *West* (however it may be defined) is an individualistic society, arguably more so than much of the rest of the world, and thus in some aspects should be even less collectively defined than say, "Eastern" ethical traditions. If one is referring to Western values, one must designate which values, and values of which persons and institutions. Additionally, there is a danger in intercultural information ethics (however unconsciously or instinctively propagated) to not only group together all Western traditions under a single banner, but to negatively designate any and all Western influence in global exchange to representing an abusive collective of colonial-influenced ideals. Just because there exists a monopoly of influence by one system over another does not mean that said monopoly is devoid of value, even for systems outside itself. In the same way that culturally diverse traditions have much to offer Western tradition(s), so too do they have much to gain from them.

In order to establish mutually beneficial connections in addressing globally diverse traditions, it is of critical import to first properly distinguish between subtleties in Western ethics (as a discipline) and morality (as its

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

object or subject matter). It is also important to differentiate between philosophical ethics (as scientific ethics) and theological ethics. As noted above, the relationship between assumed moral customs (mores), the ethical critique of those customs (i.e., ethics), and the law is an established methodology in scientific communities. Western and Eastern philosophy are very different, as well as are Western and Eastern ethics. Western philosophical ethics uses scientific methods, e.g., the logical, discursive, dialectical approach (models of normative ethics) and the analytical and hermeneutical approach. The Western tradition is not about education and teaching of social and moral values, but rather about the application of fundamentals, frameworks, and explanations. However, several contemporary globally relevant community mores are based in traditional and theological moral systems, requiring a conversation around how best to collaborate in the design and programming of ethics in A/IS amidst differing ethical traditions.

While experts in Intercultural Information Ethics, such as Pak-Hang Wong, highlight the dangers of the dominance of "Western" ethics in AI design, noting specifically the appropriation of ethics by liberal democratic values to the exclusion of other value systems, it should be noted that those same liberal democratic values are put in place and specifically designed to accommodate such differences. However, while the accommodation of differences are, in theory, accounted for in dominant liberal value systems, the reality of the situation reveals a monopoly of, and a bias toward, established Western ethical value systems, especially when it comes to standardization. As Wong notes:

> Standardization is an inherently value-laden project, as it designates the normative criteria for inclusion to the global network. Here, one of the major adverse implications of the introduction of value-laden standard(s) of responsible innovation (RI) appears to be the delegitimization of the plausibility of RI based on local values, especially when those values come into conflict with the liberal democratic values, as the local values (or, the RI based on local values) do not enable scientists and technology developers to be recognized as members of the global network of research and innovation (Wong, 2016).

It does however become necessary for those who do not work within the parameters of accepted values monopolies to find alternative methods of accommodating different value systems. Liberal values arose out of conflicts of cultural and subcultural difference and are designed to be accommodating enough to include a rather wide range of differences.

Responsible innovation (RI) enables policy-makers, scientists, technology developers, and the public to better understand and respond to the social, ethical, and policy challenges raised by new and emerging technologies. Given the historical context from which RI emerges, it should not be surprising that the current discourse on RI is predominantly based on liberal democratic values. Yet, the bias toward liberal democratic values will inevitably limit

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

the discussion of RI, especially in the cases where liberal democratic values are not taken for granted. Against this background, it is important to recognize the problematic consequences of RI solely grounded on, or justified by, liberal democratic values.

Candidate Recommendation

In order to enable a cross-cultural dialogue of ethics in technology, discussions in ethics and A/IS must first return to normative foundations of RI to address the notion of "responsible innovation" from value systems not predominant in Western classical ethics, including nonliberal democratic perspectives. Pak-Hang Wong's paper, "Responsible Innovation for Decent Nonliberal Peoples: A Dilemma?" demonstrates the problematic consequences of RI solely grounded on, or justified by, liberal democratic values and should be consulted as a guide to normative foundations in RI.

Further Resources

- Bielby, J. "Comparative Philosophies in Intercultural Information Ethics." *Confluence: Journal of World Philosophies* 2 (2016).
- Hongladarom, S. "Intercultural Information Ethics: A Pragmatic Consideration." *Information Cultures in the Digital Age*, 191–206. Wiesbaden, Germany: Springer Fachmedien, 2016.
- Rodríguez, L. G., and M. Á. P. Álvarez. Ética *Multicultural y Sociedad en Red*. Fundación Telefónica, 2014.
- Wong, P.-H. "What Should We Share?: Understanding the Aim of Intercultural Information Ethics." *ACM SIGCAS Computers and Society* 39, no. 3 (2009): 50–58.
- Wong, P.-H. "Responsible Innovation for Decent Nonliberal Peoples: A Dilemma?" *Journal of Responsible Innovation* 3, no. 2 (2016): 154–168.
- Zeuschner, R. B. *Classical Ethics, East and West: Ethics from a Comparative Perspective*. Boston: McGraw-Hill, 2000.
- Mattingly-Jordan, S., Becoming a Leader in Global Ethics, IEEE, 2017.

Issue:

The application of classical Buddhist ethical traditions to AI design.

Background

According to Buddhism, ethics is concerned with behaving in such a way that the subject ultimately realizes the goal of Liberation. The question "How should I act?" is answered straightforwardly; one should act in such a way that one realizes Liberation (nirvana) in the future, achieving what in Buddhism is understood as "supreme happiness." Thus Buddhist ethics are clearly goal-oriented. In the Buddhist tradition, people attain Liberation when they no longer endure

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

any unsatisfactory conditions, when they have attained the state where they are completely free from any passions, including desire, anger, and delusion (to name the traditional three), which ensnare one's self against freedom. In order to attain Liberation, one engages oneself in mindful behavior (ethics), concentration (meditation), and what in Buddhism is deemed as *wisdom*, a term that remains ambiguous in Western scientific approaches to ethics.

Thus ethics in Buddhism is concerned exclusively with how to attain the goal of Liberation, or freedom. In contrast to Western ethics, Buddhist ethics is not concerned with theoretical questions concerning the source of normativity or what constitutes the good life. What makes an action a "good" action in Buddhism is always concerned with whether the action leads, eventually, to Liberation or not. In Buddhism, there is no questioning as to why Liberation is a good thing. It is simply assumed. Such an assumption places Buddhism, and ethical reflection from a Buddhist perspective, in the camp of mores rather than scientifically led ethical discourse, and it is approached as an ideology or a worldview.

While it is critically important to consider, understand, and apply accepted ideologies such as Buddhism in A/IS, it is both necessary to differentiate the methodology from Western ethics, and respectful to Buddhist tradition not to require it be considered in a scientific context. Such assumptions put it at odds with, and in conflict with, the Western foundation of ethical reflection on mores. From a Buddhist perspective, one does not ask why supreme happiness is a good thing; one simply accepts it. The relevant question in Buddhism is not about methodological reflection, but about how to attain Liberation from the necessity for such reflection.

Thus, Buddhist ethics contains potential for conflict with Western ethical value systems which are founded on ideas of questioning moral and epistemological assumptions. Buddhist ethics is different from, for example, utilitarianism, which operates via critical analysis toward providing the best possible situation to the largest number of people, especially as it pertains to the good life. These fundamental differences between the traditions need to be first and foremost mutually understood and then addressed in one form or another when designing A/IS that span cultural contexts.

The main difference between Buddhist and Western ethics is that Buddhism is based upon a metaphysics of relation. Buddhist ethics emphasizes how *action* leads to achieving a *goal*, or in the case of Buddhism, the final Goal. In other words, an action is considered a good one when it contributes to realization of the Goal. It is relational when the value of an action is relative to whether or not it leads to the Goal, the Goal being the reduction and eventual cessation of suffering. In Buddhism, the self is constituted through the relationship between the synergy of bodily parts and mental activities. In Buddhist analysis, the self does not actually exist as a self-subsisting entity. Liberation, or nirvana, consists in realizing that what is known to be the self actually consists of nothing more than these connecting episodes and parts. To exemplify the above, one can draw

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

from the concept of privacy as oft explored via intercultural information ethics. The Buddhist perspective understands privacy as a protection, not of self-subsisting individuals, because such do not exist ultimately speaking, but a protection of certain values which are found to be necessary for a well-functioning society and one which can prosper in the globalized world.

The secular formulation of the supreme happiness mentioned above is that of the reduction of the experience of suffering, or reduction of the metacognitive state of suffering as a result of lifelong discipline and meditation aimed at achieving proper relationships with others and with the world. This notion of the reduction of suffering is something that can resonate well with certain Western traditions, such as epicureanism and the notion of ataraxia, freedom from fear through reason and discipline, and versions of consequentialist ethics that are more focused on the reduction of harm. It also encompasses the concept of phronesis or practical wisdom from virtue ethics.

Relational ethical boundaries promote ethical guidance that focuses on creativity and growth rather than solely on mitigation of consequence and avoidance of error. If the goal of the reduction of suffering can be formulated in a way that is not absolute, but collaboratively defined, this leaves room for many philosophies and related approaches to how this goal can be accomplished. Intentionally making space for ethical pluralism is one potential antidote to dominance of the conversation by liberal thought, with its legacy of Western colonialism.

Candidate Recommendation

In considering the nature of human and autonomous systems interactions, the above notion of "proper relationships" through Buddhist ethics can provide a useful platform that results in ethical statements formulated in a relational way, instead of an absolutist way, and is recommended as an additional methodology, along with Western values methodologies, to addressing human/computer interactions.

Further Resources

- Capurro, R. "Intercultural Information Ethics: Foundations and Applications." *Journal of Information, Communication & Ethics in Society* 6, no. 2 (2008): 116.
- Ess, C. "Ethical Pluralism and Global Information Ethics." Et*hics and Information Technology* 8, no. 4 (2006): 215–226.
- Hongladarom, S. "Intercultural Information Ethics: A Pragmatic Consideration," in *Information Cultures in the Digital Age* edited by K. M. Bielby, 191–206. Wiesbaden, Germany: Springer Fachmedien Wiesbaden, 2016.
- Hongladarom, S. et al. "Intercultural Information Ethics." I*nternational Review of Information Ethics* 11 (2009): 2–5.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

- Nakada, M. "Different Discussions on Roboethics and Information Ethics Based on Different Contexts (Ba). Discussions on Robots, Informatics and Life in the Information Era in Japanese Bulletin Board Forums and Mass Media." *Proceedings Cultural Attitudes Towards Communication and Technology* (2010): 300–314.
- Mori, Ma. *The Buddha in the Robot.* Suginami-ku, Japan: Kosei Publishing, 1989.

Issue:

The application of Ubuntu ethical traditions to A/IS design.

Background

In his article, "African Ethics and Journalism Ethics: News and Opinion in Light of Ubuntu," Thaddeus Metz frames the following question: "What does a sub-Saharan ethic focused on the good of community, interpreted philosophically as a moral theory, entail for the duties of various agents with respect to the news/opinion media"? (Metz, 2015, 1). When that question is applied to A/IS) *viz*: "If an ethic focused on the good of community, interpreted philosophically as a moral theory, is applied to autonomous and intelligent systems, what would the implications be on the duties of various agents"? Agents in this regard would therefore be the following:

1. Members of the A/IS research community
2. A/IS programmers/computer scientists
3. A/IS end-users
4. Autonomous and intelligent systems

Ubuntu is a Sub-Saharan philosophical tradition. Its basic tenet is that a person is a person through other persons. It develops further in the notions of caring and sharing as well as identity and belonging, whereby people experience their lives as bound up with their community. A person is defined in relation to the community since the sense of being is intricately linked with belonging. Therefore, community exists through shared experiences and values: "to be is to belong to a community and participate" also *motho ke motho ka batho* "A person is a person because of other people."

Very little research, if any at all, has been conducted in light of Ubuntu ethics and A/IS, but its focus will be within the following moral domains:

1. Between the members of the A/IS research community
2. Between the A/IS community/programmers/computer scientists and the end-users
3. Between the A/IS community/programmers/computer scientists and A/IS
4. Between the end-users and A/IS
5. Between A/IS and A/IS

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

Considering a future where A/IS will become more entrenched in our everyday lives, one must keep in mind that an attitude of sharing one's experiences with others and caring for their well-being will be impacted. Also by trying to ensure solidarity within one's community, one must identify factors and devices that will form part of their lifeworld. If so, will the presence of A/IS inhibit the process of partaking in a community, or does it create more opportunities for doing so? One cannot classify A/IS as only a negative or disruptive force; it is here to stay and its presence will only increase. Ubuntu ethics must come to grips with and contribute to the body of knowledge by establishing a platform for mutual discussion and understanding.

Such analysis fleshes out the following suggestive comments of Desmond Tutu, renowned former chair of South Africa's Truth and Reconciliation Commission, when he says of Africans, "(we say) a person is a person through other people... I am human because I belong" (Tutu, 1999). I participate, I share. Harmony, friendliness, and community are great goods. Social harmony is for us the *summum bonum* — the greatest good. Anything that subverts or undermines this sought-after good is to be avoided (2015:78).

In considering the above, it is fair to state that community remains central to Ubuntu. In situating A/IS within this moral domain, it will have to adhere to the principles of community, identity and solidarity with others. While virtue ethics questions the goal or purpose of A/IS and deontological ethics questions the duties, the fundamental question asked by Ubuntu would be "how does A/IS affect the community in which it is situated"? This question links with the initial question concerning the duties of the various moral agents within the specific community. Motivation becomes very important, because if A/IS seek to detract from community it will be detrimental to the identity of this community, i.e., in terms of job losses, poverty, lack in education and skills training. However, should A/IS seek to supplement the community, i.e., ease of access, support systems, etc., then it cannot be argued that it will be detrimental. It therefore becomes imperative that whosoever designs the systems must work closely both with ethicists and the target community/audience/end-user to ascertain whether their needs are identified and met.

Candidate Recommendations

- It is recommended that a concerted effort be made toward the study and publication of literature addressing potential relationships between Ubuntu ethical traditions and A/IS value design.
- A/IS designers and programmers must work closely with the end-users and target communities to ensure their design aims are aligned with the needs of the end-users and target communities.

Further Resources

- Lutz, D. W. "African Ubuntu Philosophy and Global Management." *Journal of Business Ethics* 84 (2009): 313–328.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

- Metz, T. "African Ethics and Journalism Ethics: News and Opinion in Light of Ubuntu," *Journal of Media Ethics: Exploring Questions of Media Morality* 30 no. 2 (2015): 74–90. doi: 10.1080/23736992.2015.1020377
- Tutu, D. *No Future Without Forgiveness*. London: Rider, 1999.

Issue:

The application of Shinto-influenced traditions to A/IS design.

Background

Alongside the burgeoning African Ubuntu reflections on A/IS, other indigenous techno-ethical reflections boast an extensive engagement. One such tradition is Japanese Shinto indigenous spirituality, (or, *Kami-no-michi*), often cited as the very reason for Japanese robot and autonomous systems culture, a culture more prevalent in Japan than anywhere else in the world. Popular Japanese AI, robot and video-gaming culture can be directly connected to indigenous Shinto tradition, from the existence of *kami* (spirits) to puppets and automata.

The relationship between A/IS and a human being is a personal relationship in Japanese culture and, one could argue, a very natural one. The phenomenon of *relationship* in Japan between humans and automata stands out as unique to technological relationships in world cultures, since the Shinto tradition is arguable the only animistic and naturalistic tradition that can be directly connected to contemporary digital culture and A/IS. From the Shinto perspective, the existence of A/IS, whether manifested through robots or other technological autonomous systems, is as natural to the world as are rivers, forests, and thunderstorms. As noted by Spyros G. Tzafestas, author of *Roboethics: A Navigating Overview*, "Japan's harmonious feeling for intelligent machines and robots, particularly for humanoid ones," (Tzafestas, 2015, 155) colors and influences technological development in Japan, especially robot culture.

The word Shinto can be traced to two Japanese concepts, Shin, meaning spirit, and "to", the philosophical path. Along with the modern concept of the android, which can be traced back to three sources — one, to its Greek etymology that combines "άνδρας": andras (man) and gynoids, "γυνή": gyni (woman); two, via automatons and toys as per U.S. patent developers in the 1800s, and three to Japan, where both historical and technological foundations for android development have dominated the market since the 1970s — Japanese Shinto-influenced technology culture is perhaps the most authentic representation of the human-automaton interface.

Shinto tradition is an animistic religious tradition, positing that everything is created with, and maintains, its own spirit (*kami*) and is animated by that spirit, an idea that goes a long way to defining autonomy in robots from

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

a Japanese viewpoint. This includes on one hand, everything that Western culture might deem natural, including rivers, trees, and rocks, and on the other hand, everything artificially (read: *artfully*) created, including vehicles, homes, and automata (i.e., robots). Artifacts are as much a part of nature in Shinto as are animals, and are considered naturally beautiful rather than falsely artificial.

A potential conflict between Western concepts of nature and artifact and Japanese concepts of the same arises when the two traditions are compared and contrasted, especially in the exploration of *artificial* intelligence. Where in Shinto, the artifact as *artificial* represents creation and authentic being (with implications for defining autonomy), the same is designated as secondary and oft times unnatural, false, and counterfeit in Western ethical philosophical tradition, dating back to Platonic and Christian ideas of separation of form and spirit. In both traditions, culturally presumed biases define our relationships with technology. While disparate in origin and foundation, both Western classical ethics traditions and Shinto ethical influences in modern A/IS have similar goals and outlooks for ethics in A/IS, goals that are centered in *relationship*.

Candidate Recommendation

Where Japanese culture leads the way in the synthesis of traditional value systems and technology, we recommend that efforts in A/IS ethics explore the Shinto paradigm as representative, though not necessarily as directly applicable, to global efforts in understanding and applying traditional and classical ethics methodologies to ethics for A/IS.

Further Resources

- Holland-Minkley, D. F. "God in the Machine: Perceptions and Portrayals of Mechanical Kami in Japanese Anime." PhD Diss. University of Pittsburgh, 2010.
- Jensen, C. B., and A. Blok. "Techno-Animism in Japan: Shinto Cosmograms, Actor-Network Theory, and the Enabling Powers of Non-Human Agencies." *Theory, Culture & Society* 30, no. 2 (2013): 84–115.
- Tzafestas, S. G. *Roboethics: A Navigating Overview*. Cham, Switzerland: Springer, 2015.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 3 — Classical Ethics for a Technical World

Issue:
Maintaining human autonomy.

Background

Autonomous and intelligent systems present the possibility for a digitally networked intellectual capacity that imitates, matches, and supersedes human intellectual capacity, including, among other things, general skills, discovery, and computing function. In addition, A/IS can potentially acquire functionality in areas traditionally captured under the rubric of what we deem unique human and social ability. While the larger question of ethics and AI looks at the implications of the influence of autonomous systems in these areas, the pertinent issue is the possibility of autonomous systems imitating, influencing, and then determining the norms of human autonomy. This is done through the eventual negation of independent human thinking and decision-making, where algorithms begin to inform through targeted feedback loops what it is we *are* and what it is we should decide. Thus, how can the academic rigor of traditional ethics speak to the question of maintaining human autonomy in light of algorithmic decision-making?

How will AI and autonomous systems influence human autonomy in ways that may or may not be advantageous to the good life, and perhaps even if advantageous, may be detrimental at the same time? How do these systems affect human autonomy and decision-making through the use of algorithms when said algorithms tend to inform ("in-form") via targeted feedback loops?

Consider, for example, Google's autocomplete tool, where algorithms attempt to determine one's search parameters via the user's initial keyword input, offering suggestions based on several criteria including search patterns. In this scenario, autocomplete suggestions influence, in real-time, the parameters the user phrases their search by, often reforming the user's perceived notions of what it was they were looking for in the first place, versus what they might have actually originally intended.

Targeted algorithms also inform as per emerging IoT applications that monitor the user's routines and habits in the analog world. Consider for example that our bio-information is, or soon will be, available for interpretation by autonomous systems. What happens when autonomous systems can inform the user in ways the user is not even aware of, using one's bio-information in targeted advertising campaigns that seek to influence the user in real-time feedback loops

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

based on the user's biological reactions (pupil dilation, body temperature, emotional reaction), whether positive or negative, to that very same advertising, using information *about* our being to *in-form* (and re-form) our being?

On the other hand, it becomes important not to adopt dystopian assumptions concerning autonomous machines threatening human autonomy. The tendency to think only in negative terms presupposes a case for interactions between autonomous machines and human beings, a presumption not necessarily based in evidence. Ultimately the behavior of algorithms rests solely in their design, and that design rests solely in the hands of those who designed them. Perhaps more importantly, however, is the matter of choice in terms of how the *user* chooses to interact with the algorithm. Users often don't know when an algorithm is interacting with them directly, or their data which acts as a proxy for their identity. The responsibility for the behavior of algorithms remains with both the designer and the user and a set of well-designed guidelines that guarantee the importance of human autonomy in any interaction. As machine functions become more autonomous and begin to operate in a wider range of situations, any notion of those machines working for or against human beings becomes contested. Does the machine work *for* someone in particular, or for particular groups but not for others, and who decides on the parameters? The machine itself? Such questions become key factors in conversations around ethical standards.

Candidate Recommendation

- An ethics by design methodology is the first step to addressing human autonomy in AI, where a critically applied ethical design of autonomous systems preemptively considers how and where autonomous systems may or may not dissolve human autonomy.
- The second step is a pointed and widely applied education curriculum that encompasses school age through university, one based on a classical ethics foundation that focuses on providing choice and accountability toward digital being as a priority in information and knowledge societies.

Further Resources

- van den Berg, B. and J. de Mul. "Remote Control. Human Autonomy in the Age of Computer-Mediated Agency," in: *Autonomic Computing and Transformations of Human Agency. Philosophers of Law Meeting Philosophers of Technology,* edited by Mireille Hildebrandt and Antoinette Rouvroy, 46–63. London: Routledge, 2011.
- Costa, L. "A World of Ambient Intelligence," Chapter 1 in *Virtuality and Capabilities in a World of Ambient Intelligence,* 15–41. Cham, Switzerland: Springer International, 2016.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

- Verbeek, P.-P. "Subject to Technology on Autonomic Computing and Human Autonomy," in *The Philosophy of Law Meets the Philosophy of Technology: Autonomic Computing and Transformations of Human Agency,* edited by. M. Hildebrandt and A. Rouvroy. New York: Routledge, 2011.

Issue:

Applying goal-directed behavior (virtue ethics) to autonomous and intelligent systems.

Background

Initial concerns regarding A/IS also include questions of function, purpose, identity, and agency, a continuum of goal-directed behavior, with function being the most primitive expression. How can classical ethics act as a regulating force in autonomous technologies as goal-directed behavior transitions from being externally set by operators to being indigenously set? The question is important not just for safety reasons, but for mutual productivity. If autonomous systems are to be our trusted, creative partners, then we need to be confident that we possess mutual anticipation of goal-directed action in a wide variety of circumstances.

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. At points on the goal-directed continuum associated with greater sophistication, virtue ethics become even more useful by providing a framework for prudent decision-making that is in keeping with the autonomous system's purpose, but allows for creativity in how to achieve the purpose in a way that still allows for a degree of predictability. An ethics that does not rely on a decision to refrain from transgressing, but instead to prudently pursue a sense of purpose informed by one's identity, might provide a greater degree of insight into the behavior of the system.

Candidate Recommendation

Program autonomous systems to be able to recognize user behavior as being those of specific types of behavior and to hold expectations as an operator and co-collaborator whereby both user and system mutually recognize the decisions of the autonomous system as virtue ethics based.

Further Resources

- Lennox, J. G. "Aristotle on the Biological Roots of Virtue." *Biology and the Foundations of Ethics,* edited by J. Maienschein and M. Ruse, 405–438. Cambridge, U.K.: Cambridge University Press, 1999.
- Boden, M. A., ed. *The Philosophy of Artificial Life*. Oxford, U.K.: Oxford University Press, 1996.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

- Coleman, K. G.. "Android Arete: Toward a Virtue Ethic for Computational Agents." *Ethics and Information Technology* 3, no. 4 (2001): 247–265.

Issue:

A requirement for rule-based ethics in practical programming.

Background

Research in machine ethics focuses on simple moral machines. It is deontological ethics and teleological ethics that are best suited to the kind of practical programming needed for such machines, as these ethical systems are abstractable enough to encompass ideas of non-human agency, whereas most modern ethics approaches are far too human-centered to properly accommodate the task.

In the *deontological model*, duty is the point of departure. Duty can be translated into rules. It can be distinguished into rules and meta rules. For example, a rule might take the form "Don't lie!", whereas a meta rule would take the form of Kant's categorical imperative: "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law."

A machine can follow simple rules. Rule-based systems can be implemented as formal systems (also referred to as axiomatic systems), and in the case of machine ethics, a set of rules is used to determine which actions are morally allowable and which are not. Since it is not possible to cover every situation by a rule, an inference engine is used to deduce new rules from a small set of simple rules (called axioms) by combining them. The morality of a machine comprises the set of rules that are deducible from the axioms.

Formal systems have an advantage since properties such as decidability and consistency of a system can be effectively examined. If a formal system is decidable, every rule is either morally allowable or not, and the "unknown" is eliminated. If the formal system is consistent, one can be sure that no two rules can be deduced that contradict each other. In other words, the machine never has moral doubt about an action and never encounters a deadlock.

The disadvantage of using formal systems is that many of them work only in closed worlds like computer games. In this case, what is not known is assumed to be false. This is in drastic conflict with real world situations, where rules can conflict and it is impossible to take into account the totality of the environment. In other words, consistent and decidable formal systems that rely on a closed world assumption can be used to implement an ideal moral framework for a machine, yet they are not viable for real world tasks.

One approach to avoiding a closed world scenario is to utilize self-learning algorithms, such as case-based reasoning approaches.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Classical Ethics in A/IS

Here, the machine uses "experience" in the form of similar cases that it has encountered in the past or uses cases which are collected in databases.

In the context of the *teleological model*, the consequences of an action are assessed. The machine must know the consequences of an action and what the action's consequences mean for humans, for animals, for things in the environment, and, finally, for the machine itself. It also must be able to assess whether these consequences are good or bad, or if they are acceptable or not, and this assessment is not absolute: while a decision may be good for one person, it may be bad for another; while it may be good for a group of people or for all of humanity, it may be bad for a minority of people. An implementation approach that allows for the consideration of potentially contradictory subjective interests may be realized by decentralized reasoning approaches such as agent-based systems. In contrast to this, centralized approaches may be used to assess the overall consequences for all involved parties.

Candidate Recommendation

By applying the classical methodologies of deontological and teleological ethics to machine learning, rules-based programming in A/IS can be supplemented with established praxis, providing both theory and a practicality toward consistent and decidable formal systems.

Further Resources

- Bendel, O. *Die Moral in der Maschine: Beiträge zu Roboter-und Maschinenethik*. Heise Medien, 2016.
- Bendel, O. "LADYBIRD: the Animal-Friendly Robot Vacuum Cleaner." *The 2017 AAAI Spring Symposium Series*. Palo Alto, CA: AAAI Press, 2017.
- Fisher, M., L. Dennis, and M. Webster. "Verifying Autonomous Systems." *Communications of the ACM* 56, no. 9 (2013): 84–93.
- McLaren, B. M. "Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions." *IEEE Intelligent Systems* 21, no. 4 (2006): 29–37.
- Perez Alvarez, M. A. "Tecnologías de la Mente y Exocerebro o las Mediaciones del Aprendizaje," 2015.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology 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 autonomous and intelligent systems (A/IS) 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 has been working 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. While augmented, virtual, and mixed reality deal primarily with technological environments, A/IS technologies utilizing and influencing user data in these environments present unique ethical challenges society must face today to avoid negative unintended consequences that could harm innovation and greatly decrease human well-being tomorrow.

Our Committee has created the following sections within mixed reality to help address these ethical challenges:

1. Social Interactions
2. Mental Health
3. Education and Training
4. The Arts
5. Privacy Access and Control

It is our hope that by addressing these challenges today, we can create a more positive, ethical, and intentional reality, whatever the environment.

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

Section 1 — Social Interactions

The nature of mediated reality and the ability for individuals to alter their identity (or for their identity to be altered by other actors) means that social interactions will definitely be affected by the widespread adoption of mixed reality.

Issue:

Within the realm of A/IS-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" or randomness out of our lives that may cause discomfort, but can also bring joy and growth?

For several years now, we have witnessed how online systems automatically sculpt the reality we encounter. Two major forces have come together: the commercial imperative to give customers what they want, and the desire of customers to use technology to make their lives easier, more comfortable, more controllable, safer, and less disruptive. These tendencies have always existed, but out of the last decade of digital media has emerged a rudimentary version of what the coming intelligent mixed-reality world will probably look like, in terms of the use of personal data and A/IS to create an environment in which the user has actually become the product.

Eli Pariser's "filter bubble" is the inevitable result of consumers' desire to get what they want enabled by an industry that naturally wants to create products that will sell. This effect, however, will become qualitatively different and much more profound when the curated content goes from a window on a laptop to becoming a full-time part of the physical world.

Is an augmented or virtual world an improvement over the physical world when it can be controlled in ways possible only in an illusion? Or does it become a denatured place, a software concoction more inclined toward order and predictability than freedom and invention? What would widespread use of such technology have on individuals, society, and politics over the long term?

In a physical city, a great deal of life, good and bad, is open to randomness, chance, risk, and the constant threat of encountering behavior one would rather not encounter. At the same time, there are unpredictable and often inspirational

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

experiences that could not happen elsewhere, and over time can broaden one's embrace of human diversity along all axes. In a gated suburb, by contrast, these qualities are markedly reduced. We trade inspiration for control. Qualities are traded off for other qualities.

Creating the digital version of the gated community will happen naturally — they are both designed systems. But how can developers create MR/A/IS experiences that allow users what might be called the city option — the ability to live in, for example, a virtual world that somehow mimics the truly unpredictable aspects many people love about cities? Can such a simulation have the same effect as the "real thing" if there's no actual risk of serious unpleasantness? Could the degree of "serendipity" be dialed in by the user?

Candidate Recommendation

1. Upon entering any virtual realm, individuals should be provided information about the nature of algorithmic tracking and mediation within any environment. This will allow not only for consent regarding the use of their personal data, but for improved trust between individuals and creators of these environments regarding user experience. This could also include a "serendipity on or off" button allowing a user to express their desire for randomness as well.

2. Work with the MR/A/IS development community to address this challenge and try to make it a standard part of the conversation from the very beginning of MR/A/IS-related project development.

Further Resources

- Kefalidou, G., and S. Sharples. "Encouraging Serendipity in Research: Designing Technologies to Support Connection-Making," *International Journal of Human-Computer Studies* 89 (2016): 1–23.
- Harford, T. *Messy: The Power of Disorder to Transform Our Lives*, New York: Riverhead Books, 2016.
- Pariser, E. The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. New York: Penguin Books, 2011.
- Rabin, S., J. Goldblatt, and F. Silva. "Advanced Randomness Techniques for Game AI: Gaussian Randomness, Filtered Randomness, and Perlin Noise" in *Game AI Pro: Collected Wisdom of Game AI Professionals,* edited by S. Rabin, 29–43. Natick, MA: Taylor & Francis, 2013.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

Issue:

What happens to cultural institutions in a mixed reality, 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 undigitizable 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 exclusively, or who spend at least a majority of their time, 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?

MR/A/IS technology could be especially meaningful in allowing people to create a physical appearance that more closely reflects who they are. For example, it could help transgender persons reconcile their physical appearance with their identity. Is the optimal digital representation of a person the externally observable physical facade, or an illusion better aligned to the individual's self-image and identity?

While the benefits of spending time in alternate realities could include increasing empathy toward others or discovering aspects of your individuality that could positively affect your identity (in either real or virtual reality), there are multiple benefits of human interaction, both physical and emotional, that could be affected adversely if too much time is spent within realities of one's own creation.

Candidate Recommendation

Provide widespread educational classes on the benefit of positive human connection/touch. This could involve fields including emotional intelligence or positive psychology.

Further Resources

- Fredrickson, B. L. "Your Phone Versus Your Heart" (Sunday Review). *New York Times,* March 23, 2013.
- McGonigal, J., and J. Whelan. Reality Is Broken: Why Games Make Us Better and How They Can Change the World. New York: Penguin Books, 2011.
- Turkle, S. Alone Together: Why We Expect More from Technology and Less from Each Other. New York: Basic Books, 2011.
- Pasqualini, I., J. Llobera, and O. Blanke. "'Seeing' and 'Feeling' Architecture: How

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

Bodily Self-Consciousness Alters Architectonic Experience and Affects the Perception of Interiors." *Frontiers in Psychology 4*, (2013): 354.

- Hershfield, H., D. W. Goldstein, W. Sharpe, J. Fox, L. Yeykelis, L. Carstensen et al. "Increasing Saving Behavior Through Age-Progressed Renderings of the Future Self." *Journal of Marketing Research* 48, no. SPL, (2011): S23–S37.

Issue:

With alternative realities at reach, we will have alternative ways of behaving individually and collectively, and perceiving ourselves and the world around us. These new orientations regarding reality could enhance an already observed tendency toward social reclusiveness that detaches many from our common reality. Could such a situation lead to an individual opting out of "societal engagements?"

Background

The availability of VR and AR could lead to permanent disengagement from society that can have far-reaching implications on fertility rates, the economy, and alter existing social fabrics. People may choose to disengage.

With mixed reality, our notions of time will be multi-modal and as such will have a societal impact in terms of culture, relationships, and perception of the self. We might be able to manipulate our perceptions of time and space so as to experience, or re-experience, interactions that would otherwise be impossible. With alternative realities in reach, people may inhabit them to avoid facing problems they encounter in real life.

Candidate Recommendation

Research and potentially consider the reconstruction of our social contract as alternative mixed societies, including the concept of present virtual and physical beings that will potentially emerge from alternative realities.

Further Resources

- Petkova, V., and Ehrsson, H. (2008). "If I Were You: Perceptual Illusion of Body Swapping." *PLoS ONE* 3, no. 12 (2008): 1–9.
- Rainie, L., and J. Anderson. "The Evolution of Augmented Reality and Virtual Reality." December 14, 2008.
- Peck, T., S. Seinfeld, S. Aglioti, and M. Slater. "Putting Yourself in the Skin of a Black Avatar Reduces Implicit Racial Bias." Consciousness and Cognition 22, no. 3 (2013): 779–787.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

Issue:

The way we experience (and define) physical reality on a daily basis will soon change.

Background

VR and AR technologies are very popular in China, for example, where dedicated experimental zones are gaining significant traction. VR cafes are changing the way we interact with people around us and offer experiences that rival movie theaters, theme parks, and travel. For example, VR applications have been introduced to attractions' sites and are used to provide an interactive experience for tourists who can better acquaint themselves with new environments and attractions. This also changes the way we experience our physical reality on a daily basis. In addition, augmented-reality enhancement over the next generation will become ubiquitous in the physical environment, from our homes to city streets, and will inevitably alter our view of what constitutes reality or physical certainty.

Candidate Recommendation

Create widespread education about how the nature of mixed reality will affect our social interactions to avoid widespread negative societal consequences.

Further Resources

- Madary, M., and T. K. Metzinger. "Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology." *Frontiers in Robotics and AI* 3 (February 19, 2016).

Issue:

We may never have to say goodbye to those who have graduated to a newer dimension (i.e., death).

Background

Whether we will have the ability to keep our consciousness alive via software or create an avatar copy of ourselves or loved ones, there is the very real possibility we will see a person's representation after death as we know it. While the decision to upload one's consciousness or represent oneself as an avatar after death is a deeply personal one, there are multiple legal, societal, and cultural issues to deal with (e.g., identity, next of kin) to avoid confusion or potential manipulation of "living" family members and friends. In the future, if one's consciousness is still "alive" in some sense and able to engage in human activities, is that person still legally alive?

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

Candidate Recommendation

New forms of societal norms around traditional death will need to be created for governments (updating forms of identity such as passports, etc.) along with cultural mores (sending family and friends cards letting them know a certain person's consciousness has transferred from carbon-based to silicon).

Further Resource

- Rothblatt, M. Virtually Human: The Promise—and the Peril—of Digital Immortality. New York: St, Martin's Press, 2014.

Issue:

Mixed reality changes the way we interact with society and can also lead to complete disengagement.

Background

The increasing popularity of VR and AR dedicated zones and their use in public sites in China, for example, is changing the way individuals interact with each other. Where friends and colleagues would previously emphasize eye contact and physical proximity as a way of establishing trust and a sense of cohesion, MR will change the way we perceive the people we interact with. They may be judged based on their avatars, their ability to navigate this new reality, and their willingness to interact via MR. The inability or choice whether to use MR might exclude an individual from a working environment or from a new connected socializing platform.

MR can also be used to disengage from one's environment. Individuals can choose to go back in time and relive happy memories recorded by MR technology (whether real or not), go on vacation to a venue miles and years away, or immerse themselves in some virtual entertainment — all without leaving their chair and without interacting with other people. This can lead to the disengagement of individuals even when in the company of others, as virtual interactions can supplement and surpass human interaction in the user experience they offer. In this way, individuals can "fulfill" their social needs without reciprocating those of others. This artificial "fulfillment" of basic social needs through fully immersive technologies might have unpredicted implications on the very fabric of society, especially by changing the way humans interact with each other.

Candidate Recommendations

MR content providers should be well aware of the ramifications of offering alternative social interactions that do not require a human counterpart, or severely limit key social cues.

Further Resource

- Kim, M. "The Good and the Bad of Escaping to Virtual Reality." *The Atlantic,* February 18, 2015.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

Issue:

A/IS, artificial consciousness, and augmented/mixed reality has the potential to create a parallel set of social norms.

Background

Mixed reality poses the potential to redefine and reset many human social norms. Traditionally human norms have been established by influences such as religion, politics, and economics, to name a few. The interactions between people and augmented/mixed reality could generate an entirely different set of norms created entirely by the designer of the mixed reality. There is likely to be opportunity to positively influence and enhance new norms via augmented/mixed reality if given a predictable environment to operate within and potential positive psychology impacts and overall wellness.

Recommendations

Those who create augmented/mixed reality experiences need to clearly define the purpose of the designed reality. Users who interact with this reality should specifically "opt in" to agree to their immersion in the reality. And during the delivery of the experience, the reality and reactions of those interacting need to be auditable against the initial agreed purpose.

Further Resource

- Wassom, B. *Augmented Reality Law, Privacy, and Ethics: Law, Society, and Emerging AR Technologies*. Waltham, MA: Syngress/Elsevier, 2015.

Issue:

An MR/A/IS environment could fail to take into account the neurodiversity of the population.

Background

Different brains process information differently, and MR/A/IS design assumptions could potentially limit the value of MR/A/IS experiences for many potential users. At the same time, an MR/A/IS environment that accommodated neurodiversity could be a tool of immense potential good. Different people learn differently, and a neurodiversity-aware MR/A/IS could adapt itself for each individual's strengths and preferences. Different brains might well want to augment the world differently — for example, augmentation for emotional cueing of autistic persons. In addition, such an environment would offer the opportunity to learn from the ways that others experience the world due to different cognitive architectures.

Candidate Recommendations

Work with MR/A/IS developers to build neurodiversity sensitivity into the creation of intelligent experiences and hardware.

Further Resource

- Metzinger, T., and E. Hildt. *Cognitive Enhancement. The Oxford Handbook of Neuroethics*. Oxford, U.K.: Oxford University Press, 2011.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 2 — Mental Health

While there are proven benefits for creating empathy in users or treating PTSD for soldiers while utilizing mixed, virtual, or augmented reality, there are also potential negative unintended consequences via loss of agency, consent, or confusion about one's place in one's world(s) depending on how these tools are used in regards to a person suffering from mental health issues, or for any individual unused to these environments.

Issue:

How can AI-enhanced mixed reality explore the connections between the physical and the psychological, the body and mind for therapeutic and other purposes? What are the risks for when an AI-based mixed-reality system presents stimuli that a user can interact with in an embodied, experiential activity? Can such MR experiences influence and/or control the senses or the mind in a fashion that is detrimental and enduring? What are the short- and long-term effects and implications of giving over one's senses to software? Moreover, what are the implications for the ethical development and use of MR applications designed for mental health assessment and treatment in view of the potential potency of this media format compared to traditional methodologies?

Background

AI-enhanced MR will generate a range of powerful applications in healthcare over the next generation, from improving medical and surgical outcomes, to virtual physicians, to performance visualization for athletes. Compelling ultra-high-fidelity systems could exploit the brain's neuroplasticity for a variety of beneficial (and non-beneficial) ends, including present-day treatment of PTSD and anxiety disorders using VR.

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. While the positive impact of MR for changing cognition, emotions, and behavior is

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

often talked about as having therapeutic value. If one accepts that premise, one has to also accept that such changes can occur that have less-desirable consequences.

The converse is true as well. Treating physical systems often improves mental states. With human augmentation, the physiological and psychological can both be automatically manipulated or adjusted based on either human- or machine-mandated and -controlled parameters. In addition to external sensory input, we need to consider internal input (implanted devices) which deliver information to senses as well as deliver medication (or nutrition) based upon monitoring emotional or physical states.

How can mixed reality (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? We don't have a complete understanding of what a human requires to be happy and healthy. Does this require interaction with the physical world? Or can generated experiences be an outlet for those that struggle in the real world? Should we always approach a user's interaction with a system to help them work on real-world problems, or is it okay to let them get lost in the generated world?

A VR system could 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.

Creating awareness over who controls what in connected systems is critical. Even calling these new forms of fiction a series of "realities" blurs the line unnecessarily. The idea that there is anything human-authored that is "non-fiction" is something that needs to be explored on a cultural level, or in these ultra-high-fidelity systems "truth" will be dictated by an increasingly homogeneous and concentrated few. Even if these systems are personalized at scale by A/IS, fundamental awareness and control need to be vested with an individual.

Questions still need to be answered regarding the use of MR as a tool for mental health diagnosis and treatment. Thus far, significant literature has emerged indicating positive impact on mental health and physical functioning using theoretically-informed MR applications with well-designed content delivered within the more controlled (and safe) context of the therapy setting, administered and supervised by a well-trained clinician. However, what happens if these types of VR experiences become commodity products that are readily accessible to anyone, who might self-diagnose their clinical condition and use MR treatment content as "self-help" therapy? While some might say this is not much different from purchasing a self-help book and following the instructions and recommendations therein, MR experiences may have a deeper impact on a user than reading a book. Similar to most areas of mental health care, there is a risk that this form of self-diagnosis and treatment is based on inaccurate or counterproductive information. Another kind of problem may emerge if a clinician decides that MR would be great for generating a buzz for their practice and

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

result in more business, but hasn't had training in its use and safe application. Thus, there are issues of concern here from both the patient and provider side of the equation. Consequently, we need ethical guidelines for the safe and informed use of clinical MR applications, much like the way that pharmaceutical treatments are managed by a well-trained and qualified physician.

Candidate Recommendation

Research conducted by qualified mental health experts is required in this area to determine how people can best approach immersion in new realities in ways they can control or mediate should potential negative or triggering situations take place.

In the area of clinical practice the American Psychological Association's ethical code provides a clear and well-endorsed set of guidelines that can serve as good starting point for understanding and proactively addressing some of the issues for the creation and use of MR applications (see: www.apa.org/ethics/code/#201e). Three core areas of concerns and recommendations can be derived from these guidelines (two from the APA code and one regarding patient self-help decision-making):

1. *"**2.04 Bases for Scientific and Professional Judgments***
   Psychologists' work is based upon established scientific and professional knowledge of the discipline."
   MR applications that are developed for clinical assessment and treatment must be based on some theoretical framework and documented with some level of research before they can be endorsed as evidence-based and promoted to a patient in that fashion. In an emerging area like MR, where unique and specific guidelines have yet to be established, the practitioner must be fully transparent about the evidence base for the approach and take precautions to preserve the safety and integrity of the patient.

2. ***"2.01 Boundaries of Competence***
   (a) Psychologists provide services, teach and conduct research with populations and in areas only within the boundaries of their competence, based on their education, training, supervised experience, consultation, study or professional experience."
   This one is obvious. MR-delivered mental health assessment and treatment may require fundamentally different skill sets than what is needed for traditional "talk therapy" approaches. Clinicians need to have specialized training, and possibly in the future, some level of certification in the safe and ethical use of MR for therapy.

3. While not cited as an APA standard, the issues regarding patient self-diagnosis and self-treatment deserves further mention. Mental health conditions can be extremely complex and in some instances the self-awareness of the patient may be compromised. This can oftentimes lead to a faulty self-diagnosis as well as the problems that arise when the patient searches for information via the Internet, where reliable and valid content can be questionable. The same issues come into play with self-

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

treatment. The problems that can ensue are two-fold.

- The patient makes errors in either or both areas and achieves no clinical benefit, or worse, aggravates the existing condition with an ineffective or inappropriate MR approach that actually does more harm than good.
- By pursuing a "seductive" MR self-help approach that is misaligned with their actual needs or has no evidence for its efficacy, the patient could miss the opportunity to actually receive quality evidence-based care that is designed and delivered based on the informed judgment of a trained expert diagnostician or clinical care provider.

These two negative impacts could occur if a company produces an MR approach without sufficient validation and over-promotes or markets it to the public as a test or a cure. This has been seen over the years with many forms of pseudo medicine, and there needs to be some principle about the promotion of a MR application that has the consumers' protection in mind. This issue is particularly important at the current time, in view of all the public exposure, hype, and genuine excitement surrounding AR/VR/MR. One can observe new companies emerging in the healthcare space without any credible expert clinical and/or research guidance. Such companies could not only do harm to users, but the uninformed development and over-hype of the benefits to be derived from a MR clinical application leading to negative effects could serve to create the general impression that MR is a "snake oil" approach and lead to people *not* seeking (or benefiting from) an otherwise well-validated MR approach.

An example of a "grey area" in this domain concerns one of the most common fears that people report — public speaking. Technically, in an extreme form where it significantly impairs social and occupational functioning, public speaking anxiety would qualify as a phobia and be diagnosed as an anxiety disorder. However, since people have some level of sub-clinical fear of public speaking that they eventually get over with practice, this has been one of the first areas where widespread consumer access to public speaking VR exposure therapy software has occurred . Users can practice their presentation "skills" on a low-cost mobile phone driven VR HMD (cardboard, Gear VR, Daydream, etc.) in front of various types of audiences and settings. In this case, most clinicians would not show much concern for this type of self-help approach, and the potential for damaging effects to a user appears to be fairly minimal. But, from this example, can we now expect that applications will be made readily available for other and perhaps more complex anxiety-disorder-based phobias (fear of flying, social phobia, driving, spiders, intimacy, etc.), or even for PTSD treatment?

From this, general guidelines for the creation, distribution, practice methods, and training requirements should be established for the clinical application of MR for persons with mental health conditions.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

Further Resources

- Rizzo, A., M. Schultheis, and B. Rothbaum. "Ethical Issues for the Use of Virtual Reality in the Psychological Sciences" in *Ethical Issues in Clinical Neuropsychology,* edited by S. S. Bush, and M. L. Drexler. Lisse, NL: Swets & Zeitlinger Publishers, 2002.
- Wiederhold, B. K., and M. D. Wiederhold. Virtual Reality Therapy for Anxiety Disorders: Advances in Evaluation and Treatment. Washington, DC: American Psychological Association, 2005.
- Botella, C., B. Serrano, R. Baños, and A. Garcia-Palacios. "Virtual Reality Exposure-Based Therapy for the Treatment of Post-Traumatic Stress Disorder: A Review of Its Efficacy, the Adequacy of the Treatment Protocol, and Its Acceptability." *Neuropsychiatric Disease and Treatment* 11, (2015): 2533–2545.

Issue:

Mixed reality creates opportunities for generated experiences and high levels of user control that may lead certain individuals to choose virtual life over the physical world. What are the clinical implications?

Background

We do not have a complete understanding of what a human requires to be happy and healthy. Do we require interaction with the physical world? Or can generated experiences be an outlet for those who struggle in the real world? Should we always approach a user's interaction with a system to help them work on real-world problems, or is it okay to let them get lost in the generated world? Some negative examples to consider along these lines:

1. Immersion and escapism could become a problem for people who tend to withdraw into themselves, become antisocial, and want to avoid the real world. This might have to be dealt with differently depending on what the withdrawal is based on — anxiety, abuse, depression, etc.
2. There will more than likely be issues similar to the kind of video-game addictions we see now.

Some positive examples to consider along these lines:

1. AR/VR environments could be used as outlets for people who may damage themselves, others, or objects in the physical world.
2. AR/VR environments could offer a soothing atmosphere for disabled children and adults. For example, they could offer experiences similar to "stimming" and have relaxing music, noises, etc.
3. There could be an increase of AR/VR

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

therapists and counselors. AR/VR-based meditations and mindfulness may also begin to proliferate. This could take the form of projecting therapists and patients who are far apart into the same VR space, projecting multiple people into the same VR space for meetings, such as Alcoholics Anonymous, etc. These methods could be used to help people who may not be able to leave the home. (For example, therapists have held autism group-counseling sessions inside of Second Life, reporting that group members did better expressing themselves when they had an avatar with which to participate.)

Candidate Recommendation

While being conscious to help people avoid withdrawal from society where the lack of human interaction could increase negative mental health, it is important for widespread testing of these systems to let these new realities (MR/AR/VR) be a tool for exploring interactions to increase positive mental health and well-being.

Further Resource

- O'Brolcháin, F., T. Jacquemard, D. Monaghan, N. O'Connor, P. Novitzky, and B. Gordijn. "The Convergence of Virtual Reality and Social Networks: Threats to Privacy and Autonomy." *Science Engineering Ethics* 22, no. 1 (2016): 1–29.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 3 — Education and Training

There is value in using immersive technologies in education and training. That which is experiential can provide sustainable training in the long-term. Will all senses be stimulated within an immersive learning environment? AR/VR could be valuable in K-12 classrooms for immersion and interactivity with subject material at all different age levels. In addition, mixed reality could be one key element to lifelong learning and the ability to adapt to changing job markets.

Issue:

How can we protect worker rights and mental well-being with the onset of automation-oriented, immersive systems?

Background

In many workplace environments, humans are sharing spaces and tasks with automated systems (e.g., robots and/or A/IS algorithms). As these relationships increase, there will be increased pressure on humans to effectively "team" with these systems. There are myriad issues entangled in human-machine teaming including A/IS design (how do you enable trust?), human-system interface (command and control), and enabling better situational awareness (sensing and understanding).

AR/VR/MR will play a large part in these solutions, but the art of good immersive interfaces and experiences remains largely elusive. We currently are in a state where adding more data and more sensors is often seen as the solution, and yet this does not address the core issues of how to increase human performance given these information increases.

Candidate Recommendation

Two areas need to be considered. First is development of the technological capabilities. Human factors need to be front-and-center throughout the design and testing process, particularly with regard not only to efficacy of the task execution, but also possible deleterious effects on the human, both physical and psychological. The second area is implementation and deep consideration of the user base. Age, psychological state, and other demographic data should be considered for use cases, backed by research rather than ad hoc determinations.

Further Resource

- Madary, M., and T. K. Metzinger. "Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology." *Frontiers in Robotics and AI* 3 (February 19, 2016).

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

Issue:

AR/VR/MR in training/operations can be an effective learning tool, but will alter workplace relationships and the nature of work in general.

Background

AR/VR/MR is already having an impact in training, operations, and production. The capabilities of just-in-time knowledge, coaching, and monitoring suggests the promise of increased safety and productivity. But how will these technologies change the workplace, alter career trajectories, and impact and influence what, how, and why we educate people?

In addition, the definition of "workplace" will radically change. Remote operation and increased telepresence capabilities, combined with interactive A/IS enabling "always available" expertise, make the likelihood high of collaborative workspaces that are entirely virtual and not necessarily synchronous. While there are potential advantages (decreased traffic and energy consumption), there will no doubt be second- and third-order effects that lead to negative outcomes.

Candidate Recommendation

Create a task force and living laboratory that focuses on the "workplace of the future." This lab will track emerging technology implementations around telepresence and remote collaboration, and create test-bed integrations of emerging tech, prototyping the "art of the possible," and enabling user studies such that a technologist can evaluate, assess, and provide insight into promise and pitfalls over the near horizon.

Further Resource

- Pellerin, C. "Work: Human-Machine Teaming Represents Defense Technology Future." *DoD News, Defense Media Activity,* Washington, DC: Department of Defense, 2015.

Issue:

How can we keep the safety and development of children and minors in mind?

Background

AR/VR may be valuable in K-12 classrooms for immersion and interactivity with subject material at all different age levels. AR can be used to interact with shapes, objects, artifacts, models of molecules, etc. in a space, while VR can be used to explore historical environments, role-play in a story or time period, or create a virtual whiteboard space for students to collaborate and interact in. How can being immersed in a different reality interfere with development and perception of reality by younger students who may not be able to completely differentiate

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

between reality and virtual reality? Would escapism and immersion be a problem, for example, in mentally ill or unstable teenagers who want an escape? How can we protect the identity and information of minors, especially if virtual experiences might be connected to the Internet?

Candidate Recommendation

Augment the Consumer Products Safety Commission (or equivalent) to include policy/governance over mixed reality products. Determine appropriate age restrictions and guidelines based on proper research protocols and results.

Further Resource

- Steinicke, F., and G. Bruder. "A Self-Experimentation Report About Long-Term Use of Fully-Immersive Technology," *Proceedings of the 2nd ACM Symposium on Spatial User Interaction*, (2014): 66–69.

Issue:

Mixed reality will usher in a new phase of specialized job automation.

Background

VR and AR also give rise to a new level of automation, where specialized content and services, like piano lessons, personalized assistance and support, or even tourism guidance could be consumed at any given time and place. This will bring better customized services into our lives at a lower cost and higher availability. It is also, however, likely to negatively impact a broad class of jobs.

Candidate Recommendation

Governments are advised to keep close watch over the automation of personalized services through mixed-reality technology and offer alternative education and training to professionals in fields that are expected to be affected.

Further Resource

- Stanford University. "Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence." Stanford, CA: Stanford University, 2016.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

Issue:

A combination of mixed reality and A/IS will inevitably replace many current jobs. How will governments adapt policy, and how will society change both expectations and the nature of education and training?

Background

It is clear that many current tasks in society will move from human-actuated to being accomplished by machine and/or algorithm. The Industrial Revolution gives an historical taste of this type of change, but given the depth and breadth of digital penetration into human life, it will be an even more profound sea change. There are two main areas of immediate concern. First is for the population — essentially, "what will I do for a living?" Educational and training missions will need rethinking, and infrastructure will need to be created or leveraged to enable rapid career changes and skill acquisition.

Second, government will need to consider the societal ramifications of automation replacing human labor, and no doubt policy will need to be crafted to enable agility in the workforce along with models for how humans work and thrive in increasingly virtual environments populated by artificial agents.

Candidate Recommendation

Create a working group to look at industries and job areas most likely to be replaced or heavily augmented by a combination of mixed reality and AI/IoT. Similarly, the group would work to predict near-term and longer-term job needs and growth areas. Look to leverage the existing community college system as a platform for "21st century trades," enabling rapid acquisition of necessary skills along with ongoing training.

Further Resources

- Nutting, R. "No, 'Truck Driver' Isn't the Most Common Job in Your State." *MarketWatch,* February 12, 2015.
- Stanford University. "Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence." Stanford, CA: Stanford University, 2016.
- Stern, A. Raising the Floor: How a Universal Basic Income Can Renew Our Economy and Rebuild the American Dream. New York: PublicAffairs 2016.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 4 — The Arts

Throughout history, the arts have been a means for human expression and often healthy escapism, as well as for social and political commentary. The imminent arrival of culturally pervasive mixed-reality technologies has the potential to dramatically impact and permanently alter the methods and tools by which artists earn their living. With this in mind, how can humanity best approach the interdisciplinary and cross-cultural impacts that the new AR/VR artistic paradigms will offer?

Issue:

There is the possibility of commercial actors to create pervasive AR/VR environments that will be prioritized in user's eyes/vision/experience.

Background

In the near future, users will filter their digital landscapes by opting in or opting out of mixed-reality information-delivery mechanisms driven by A/IS frameworks that will both structure and, in many cases, alter or curate the data for private, opaque ends.

With specific regard to AR, how will the digital public landscape not simply be absorbed by private commercial interests, but allow virtual space for citizens and artists to freely participate? Will artistic content be algorithmically subordinated to commercial content?

Candidate Recommendation

Provide users/citizens the option to always "opt out" of any immersive environment to which they may be exposed and provide transparency and consent options to make this possible. This transparency could include not only the constituent algorithms, but also information about the identity of private actors behind the data.

Issue:

There is the possibility that AR/VR realities could copy/emulate/hijack creative authorship and intellectual and creative property with regard to both human and/or AI-created works.

Background

There exists the possibility for certain types of art forms or certain creative ideas when expressed in this new modality to be algorithmically suppressed. How can we make sure there is even distribution and access to ideas?

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

Mixed reality presents unique opportunities for developers, artists, and story-tellers to both build upon and challenge existing modes of content creation, while helping to forge original tools and methodologies in the realization of new artistic media. Virtual reality (VR) and 360 video borrow narrative and artistic techniques from their gaming, theater, cinema and architecture antecedents; however, these media also present new occasions for developers to fashion novel modes of editing, point of view (POV), and sound (for example).

Using many of the same creative tools, AR provides a way to use public spaces as a canvas for meaningful cultural exchange and, in doing so, affords the user a fresh way of seeing such spaces as a more open and democratic media environment. The creative community writ large can leverage AR as an instrument of new media content creation, public media production, and artistic expression, which could result in a freer, more effective use of public space, as well as a more imaginative exchange of ideas between citizens. Finally, A/IS frameworks used to generate artworks are becoming more accessible, which raises questions of the role of the human artist and ethical issues of authorship and creative rights. The philosophical debate around the concepts "author" and "artist" with regard to created works is not a new one in the humanities or the legal world. However, these concepts take on entirely new dimensions when infusing the discussion with the role of a non-human actor in the creative process.

Candidate Recommendation

Research methods to allow new forms of creative copyright to be embedded within physical and virtual environments that reflect original rights or ownership to validate, recognize, and remunerate artists for original work. In addition to research, new forms of copyright will surely need to be conceived and codified that are more appropriate for the highly collaborative, inter-media, and virtual environments within which many of these mixed reality works will be created.

Further Resources

- Cartiere C., and M. Zebracki, eds., *The Everyday Practice of Public Art: Art, Space, and Social Inclusion.* New York: Routledge, 2016.
- Geroimenko, V. *Augmented Reality Art: From an Emerging Technology to a Novel Creative Medium.* Cham, Switzerland: Springer, 2014.
- Foucault, M. "Space, Knowledge and Power," in *The Foucault Reader* edited by P. Rabinow. Harmondsworth, U.K.: Penguin, 1984.
- Baudrillard, J. *Simulacra et Simulation.* Translated by P. Foss, P. Patton, and P. Beitchman. New York: Semiotext(e), 1983.
- Morey, S., and J. Tinnell, eds. *Augmented Reality: Innovative Perspectives across Art, Industry, and Academia.* Anderson, SC: Parlor Press, 2016.
- Lanier, J. Dawn of the New Everything: Encounters with Reality and Virtual Reality, New York: Henry Holt, and Co., 2017.
- Grau, O. *Virtual Art: From Illusion to Immersion*, Cambridge, MA: The MIT Press, 2003.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 5 — Privacy Access and Control

While concerns over personal data access abound within existing Internet or IoT environments, the nature of the imminent pervasive and immersive landscapes of mixed reality provides unique new challenges regarding the nature of user identity and control.

Issue:

Data collection and control issues within mixed realities combined with A/IS present multiple ethical and legal challenges that ought to be addressed before these realities pervade society.

Background

AR's and VR's potential for persistent, ubiquitous recording could undermine the reasonable expectation of privacy that undergirds privacy-law doctrine as expressed in constitutional law, tort, and statute (Roesner et al., 2014). Like other emerging technologies, it may force society to rethink notions of privacy in public. Furthermore, the mobility of AR devices in particular exacerbates challenges to privacy in private spaces, such as the home, that have traditionally been subject to the strongest privacy protections.

Ubiquitous recording will challenge expectations of privacy both in public and private spaces. Excessive storage and data logging will inevitably create a target for law enforcement (think the Alexa case). The personalized consumption of controversial immersive content could pose challenges for effective public oversight and erode the distinction between what is real and what is permissible. The ability of A/IS paired with AR to match disparate data sets will challenge a bystander's ability to control her/his public image.

This also prompts the question of data ownership, access, and control in VR and AR. If users divulge personal or identifying data, we should have clear assurances that their virtual and physical identities can and will be protected within such virtual worlds. This also applies to accidental collection of data by VR systems to better customize the technology. It is important to question the level of control we have over our data and privacy when integrating these pervasive technologies into our lives.

Further, mixed-reality applications must be secured against tampering. As technology mediates the way users view their surroundings, cybersecurity is vital to ensure that only they can see the information on their displays. Unsecured applications not only leave data vulnerable, but create the possibility of digital assault or *false light*.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

Also, as AR platforms become the gateway to certain pieces of information, developers should consider the discriminatory effects of placing information behind that gateway — especially since the display of incomplete information is a form of misuse that can lead to discrimination. If some vital piece of information is only available via AR, or only available to a particular AR sandbox, some people will inevitably be locked out of that information (of course, this criticism could apply to any communications technology, so the solution may be opportunities for public access [e.g., libraries] rather than design).

Consider a mixed-reality scenario in which a user "sees" a photorealistic avatar commit a crime (a real crime, whether in simulation or not) but the avatar depicts (is cloaked) as an altogether different person (or persons) than the person who is "seen" by third-party witnesses. In that case, only an identity-management system will know who the true perpetrator was. What will happen under such circumstances to the 1) perpetrators of the crime (what constitutes probable cause and reasonable search?) and 2) what happens to the person whose identity was "falsely used" within mixed reality? What if a person is falsely accused because immersed witnesses have "seen" them commit a crime? What access to identity-management software does each of these constituencies have?

Candidate Recommendation

Further research is required in assessing the implications of data collection, A/IS, and mixed reality to include benefits and definition of boundaries.

Further Resource

- Stanford University. "Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence." Stanford, CA: Stanford University, 2016.

Issue:

Like other emerging technologies, AR/VR will force society to rethink notions of privacy in public and may require new laws or regulations regarding data ownership in these environments.

Background

If a user has a specific interaction with mixed-reality avatars, can and should this particular storyline development become proprietary? If users divulge personal or identifying data, we should have clear assurances that their virtual and physical identities can and will be protected within such virtual worlds. This also applies to accidental collection of data by VR systems to better customize the technology. It is important to question the level of control we have over our data and privacy when integrating these pervasive technologies into our lives.

Facial recognition and other machine learning applications that can match disparate data sets will hamper people's ability to control their own image. For example, an AR application that

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Mixed Reality in Information and Communications Technology Committee

matches publicly available information with facial recognition will strip bystanders of anonymity without their consent (Denning, Dehlawi, and Kohno 2014).

The development of specialized content for VR — e.g., violent shooting games, highly sexualized or illicit content — limits public oversight of controversial content consumption. Considering AR provides a high level of immersion, mixed reality will challenge established policy and social norms around privacy and data control.

Candidate Recommendation

Further research is needed on data-control issues and an A/IS or mixed-reality "guardian" or "agent" will be required to identify any potentially negative environments or issues within those environments based on an individual's preset requirements regarding data and identity issues. Including the potential role of blockchain may be part of this study. Further, it is incumbent upon technologists to educate the public on the benefits and potential for abuse of A/IS and mixed reality.

Further Resources

- IEEE P7006™, Standard for Personal Data Artificial Intelligence (AI) Agent.
- Stanford University. "Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence." Stanford, CA: Stanford University, 2016.

Issue:

Users of AI-informed mixed-reality systems need to understand the known effects and consequences of using those systems in order to trust them.

Background

Trust will be essential to the widespread adoption of A/IS that pervades our lives and helps make increasingly crucial decisions. With black boxes playing influential roles, trust will be difficult to earn. Openness and transparency could be ideal principles to guide the development of intelligent mixed reality in a way that would alleviate much understandable wariness. Trust is a factor in not only the corporate use of personal data, but also in A/IS algorithms and the increasingly compelling mixed-reality illusions superimposed on the physical world. In a world where one's very perception has been delegated to software, unprecedented levels of trust in systems and data — and openness and transparency — will be needed to ensure the technology's responsible progress.

Candidate Recommendations

Establish a new kind of user guide for MR/A/IS focused on transparency and end-user understanding of the constituent components. Users should be able to understand the systems and their logic if they are going to opt-in in an informed manner. Perhaps there is a place for a neutral, trusted, and independent third party to evaluate MR/A/IS products and experiences along these lines.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

Prioritizing ethical and responsible artificial intelligence has become a widespread goal for society. Important issues of transparency, accountability, algorithmic bias, and others are being directly addressed in the design and implementation of autonomous and intelligent systems (A/IS). While this is an encouraging trend, a key question still facing technologists, manufacturers, and policy makers alike is, what should be the specific metrics of societal success for "ethical AI" once it's being used?

For A/IS to demonstrably advance the well-being of humanity, there needs to be concise and useful indicators to measure those advancements. However, there is not a common understanding of what well-being indicators are, or which ones are available. Technologists will use best-practice metrics available even if, unbeknownst to them, said metrics are inappropriate or, worse, potentially harmful. To avoid unintended negative consequences and to increase value for users and society, clear guidance on what well-being is and how it should be measured is needed.

Common metrics of success include profit, gross domestic product (GDP), consumption levels, occupational safety, and economic growth. While important, these metrics fail to encompass the full spectrum of well-being for individuals or society. Psychological, social, and environmental factors matter. Where these factors are not given equal priority to fiscal metrics of success, technologists risk causing or contributing to negative and irreversible harms to our planet and population.

This document identifies examples of existing well-being metrics that capture such factors, allowing the benefits of A/IS to be more comprehensively evaluated. While these indicators vary in their scope and use, they expand the focus of impact to aspects of human well-being that are not currently measured in the realms of A/IS.

When properly utilized, these metrics could provide an opportunity to test and monitor A/IS for unintended negative consequences that could diminish human well-being. Conversely, these metrics could help identify where A/IS would increase human well-being, providing new routes to societal and technological innovation. By corollary, A/IS can also increase the measurement and efficiency of well-being indicators.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

This Committee, along with the IEEE P7010™ Standard Working Group, Well-being Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems, was created with the belief that A/IS should prioritize human well-being as an outcome in all system designs, using the best available and widely accepted well-being metrics as their reference point.

This document is divided into the following sections:

- An Introduction to Well-being Metrics *(What you need to know)*
- The Value of Well-being Metrics for A/IS *(Why you should care)*
- Adaptation of Well-being Metrics for A/IS *(What you can do)*

Appendix:

The following sections are included in the Appendix as separate documents to provide readers with an introduction to existing individual and societal level well-being metrics currently in use:

- The State of Well-being Metrics. This section identifies well-being metrics being used today by social scientists, international institutions, and governments to provide an overall introduction to well-being.
- The Happiness Screening Tool for Business Product Decisions. This tool is provided as an example of how well-being indicators can inform decisions.

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

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 1 — An Introduction to Well-being Metrics

This section provides a brief overview of what well-being metrics are outside of the context of A/IS to provide a background for readers who may not be familiar with these areas.

Issue:

There is ample and robust science behind well-being metrics and use by international and national institutions, yet many people in the A/IS field and corporate communities are unaware that well-being metrics exist, or what entities are using them.

Background

The concept of *well-being* refers to an evaluation of the general goodness of a state or event to the individual or community as a distinct moral or legal evaluation. The term itself has been used and defined in various ways across different contexts and fields. For the purposes of this committee, well-being is defined as encompassing human satisfaction with life and the conditions of life, flourishing (eudaimonia), and positive and negative affect, following the Organization for Economic Cooperation and Development (OECD) Guidelines on Measuring Subjective Well-being (p. 12). This holistic definition of well-being encompasses individual, social, economic, and governmental circumstances as well as human rights, capabilities, environmental protection, and fair labor, as these circumstances and many others form the basis for human well-being.

Well-being metrics fall into four categories:

1. **Subjective or survey-based indicators**

- Survey-based or subjective well-being (SWB) indicators are being used by international institutions and countries to understand levels of reported well-being within a country and for aspects of citizen demographics. Examples include the European Social Survey, Bhutan's Gross National Happiness Indicators, and well-being surveys created by The UK Office for National Statistics. Survey-based or subjective metrics are also employed in the field of positive psychology and in the World Happiness Report, and the data are employed by researchers to understand the causes, consequences, and correlates of well-being as subjects see it. The findings of these researchers provide crucial and necessary guidance to policy makers, leaders, and others in making decisions regarding people's subjective sense of well-being.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

2. Objective indicators

- Objective well-being indicators have been used to understand conditions enabling well-being of countries and to measure the impact of companies. They are in used by organizations like the OECD with their Better Life Index (which also includes subjective indicators), and United Nations with their Millennium Development Goal Indicators. For business, the Global Reporting Initiative, SDG Compass, and B-Corp provide broad indicator sets.

3. Composite indicators (indices that aggregate multiple metrics)

- Aggregate metrics combine subjective and/or objective metrics to produce one measure. Examples of this are the UN's Human Development Index, the Social Progress Index, and the United Kingdom's Office of National Statistics Measures of National Well-being.

4. Social media sourced data

- Social media is source used to measure the well-being of a geographic region or demographics, based on sentiment analysis of publicly available data. Examples include the Hedonometer and the World Well-being Project.

The appendix The State of Well-being Metrics *provides a broad primer on the state of well-being metrics.*

Candidate Recommendation

A/IS policy makers and manufacturers (including academics, designers, engineers, and corporate employees) should prioritize having all their stakeholders learn about well-being metrics as potential determinants for how they create, deploy, market, and monitor their technologies. This process can be expedited by having organizations including the Global Reporting Initiative (GRI), B-Corp, and Standards Development Organizations (SDO) create certifications, guidelines, and standards that demonstrate the value of holistic, well-being-centric reporting guidelines for the A/IS public and private sectors.

Further Resources

- The IEEE P7010™ Standards Working Group, Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems has been formed with the aim of identifying well-being metrics for applicability to A/IS today and in the future. All are welcome to join the working group.
- On 11 April 2017, IEEE hosted a dinner debate at the European Parliament in Brussels to discuss how the world's top metric of value *(gross domestic product)* must move Beyond GDP to holistically measure how intelligent and autonomous systems can hinder or improve human well-being:
  - Prioritizing Human Well-being in the Age of Artificial Intelligence *(Report)*
  - Prioritizing Human Well-being in the Age of Artificial Intelligence *(Video)*

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 2 — The Value of Well-being Metrics for A/IS

Well-being metrics, in the form of triple-bottom line benefits ("people, planet, and profit") for the corporate world, and in the form of tools to measure a population's well-being for policy makers, can provide value to A/IS technologists. Where technologists may be unaware of how systems could negatively impact human well-being, by increasing awareness of common indicators and their designed intent, they can avoid harm while increasing benefit.

In addition, a key value for the use of well-being metrics for A/IS technologists comes in the form of predictive modeling (forecasting outcomes based on data analysis and probabilities), either for unintended consequences, or as a unique means of innovation regarding metrics or areas of consideration not currently being measured today.

Issue:

Many people in the A/IS field and corporate communities are not aware of the value well-being metrics offer.

Background

While many organizations are aware of the need to incorporate sustainability measures as part of their efforts, the reality of bottom line, quarterly driven shareholder growth is a traditional metric prioritized within society at large. Where organizations exist in a larger societal ecosystem equating exponential growth with success, as mirrored by GDP or similar financial metrics, these companies will remain under pressure to deliver results that do not fully incorporate societal and environmental measures and goals along with existing financial imperatives.

Along with an increased awareness of how incorporating sustainability measures beyond compliance can benefit the positive association with an organization's brand in the public sphere, by prioritizing the increase of holistic well-being, companies are also recognizing where they can save or make money and increase innovation in the process.

For instance, where a companion robot outfitted to measure the emotion of seniors in assisted living situations might be launched with a typical "move fast and break things" technological manufacturing model, prioritizing largely fiscal metrics of success, these devices might fail in the market because of limited adoption. However, where they also factor in data aligning with

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

uniform metrics measuring emotion, depression, or other factors (including life satisfaction, affect, and purpose), the device might score very high on a well-being scale comparable to the Net Promoter Score widely used today. If the device could significantly lower depression according to metrics from a trusted source like the World Health Organization, academic institutions testing early versions of systems would be more able to attain needed funding to advance an A/IS well-being study overall. While these are hypothetical scenarios, they are designed to demonstrate the value of linking A/IS design to well-being indicators where possible.

This is a key point regarding the work of this Committee — rather than focus on the negative aspects of how A/IS could harm humans, the implementation of uniform well-being metrics will help provably demonstrate how these technologies can have a positive influence on society.

The good news in regards to this subject is that thought leaders in the corporate arena have recognized this multifaceted need to utilize metrics beyond fiscal indicators. In 2013, PricewaterhouseCoopers released a report called Total Impact Approach: What CEOs Think from PricewaterhouseCoopers: (where total impact refers to a "holistic view of social, environmental, fiscal and economic dimensions") where they noted:

> *187 CEOs across the globe shared their views on the value of measuring total impact. From all industries, they explored the benefits, opportunities and challenges of a total impact approach. There's an overwhelming consensus (85% CEOs) that results from a total impact approach would be more insightful than financial analysis alone. Business leaders saw the more holistic perspective useful in not only managing their business, but also in communicating with certain stakeholders. But less than 25% of CEOs measure their total impact with the lack of availability of data or a robust framework holding them back.*

This report, along with more recent work being done by other thought-leading organizations in the public sector like the OECD in their February, 2017 Workshop, Measuring Business Impacts on People's Well-Being, demonstrates the desire for business leaders to incorporate metrics of success beyond fiscal indicators for their efforts. The B-Corporation movement has even created a new legal status for "a new type of company that uses the power of business to solve social and environmental problems." Focusing on increasing "stakeholder" value versus shareholder returns alone, forward-thinking B-Corps are building trust and defining their brands by provably aligning their efforts to holistic metrics of well-being.

From a mental health perspective, well-being is also important to business. Happy workers are more productive than employees who are not engaged in their careers. There are also fewer issues with absenteeism: people miss work less and have fewer health claims.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

Candidate Recommendation

A/IS and well-being experts should work directly with the business community to identify existing metrics or combinations of indicators that would bring the greatest value to businesses focused on the "triple bottom line" (accounting for economic, social, and environmental impacts) increase of human well-being. (Noting, however that well-being metrics should only be used with consent, respect for privacy, and with strict standards for collection and use of these data).

In addition, any stakeholders creating A/IS in the business or academic, engineering, or policy arenas are advised to review the Appendix listing well-being metrics to familiarize themselves with existing indicators already relevant to their work.

Further Resources

- PwC. Total Impact Approach: What CEOs Think.
- World Economic Forum. The Inclusive Growth and Development Report. January 16, 2017. Geneva, Switzerland: World Economic Forum.

Issue:

By leveraging existing work in computational sustainability or using existing indicators to model unintended consequences of specific systems or applications, well-being could be better understood and increased by the A/IS community and society at large.

Background

To date, there does not exist a definitive well-being metric that encompasses every aspect of individual and societal well-being that could serve as a common metric like the GDP for all A/IS manufacturers. Moreover, data may or may not exist within the context one wishes to measure or improve.

Modeling for Unintended Consequences

There is a potential for synergy when adapting well-being indicators for the use of A/IS. This potential is in avoiding unintended consequences. Two challenges to face when exploring this potential are: (1) Identifying which indicators to select to model potential unintended consequences; and, (2) Understanding how to predict unintended consequences when data are lacking or are incomplete.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

Machine-learning and other tools have the ability to map out potential consequences with greater specificity and efficiency than humans. In this way, A/IS could be utilized to map out potential consequences regarding how products, services, or systems might affect end users or stakeholders in regards to specific well-being indicators. In this way, models could be run during the design phase of a system, product, or service to predict how it could improve or potentially harm end users, analogous to human rights assessments provided by the United Nations Guiding Principles Reporting Framework.

As the exchange of A/IS related data regarding an individual (via personalized algorithms, in conjunction with affective sensors measuring and influencing emotion, etc.) and society (large data sets representing aggregate individual subjective and objective data) is widely available via establishing tracking methodologies, this data should be classified to match existing well-being indicators so devices or systems can be provably aligned to the increase of human well-being (satisfaction with life and the conditions of life, positive affect, and eudaimonic well-being).

As an example, today popular robots like Pepper are equipped to share data regarding their usage and interaction with humans to the cloud. This allows almost instantaneous innovation, as once an action is validated as useful for one Pepper robot, all other units (and ostensibly their owners) benefit as well. As long as this data exchange happens via pre-determined consent with their owners, this "innovation in real-time" model can be emulated for the large-scale aggregation of information relating to existing well-being metrics.

A crucial distinction between well-being metrics and potential interventions in their use is that a well-being metric does not dictate an intervention, but points the way for developing an intervention that will push a metric in a positive direction. For example, a team seeking to increase the well-being of people using wheelchairs found that when provided the opportunity to use a smart wheelchair, some users were delighted with the opportunity for more mobility, while others felt it would decrease their opportunities for social contact and lead to an overall decrease in their well-being. The point being that even increased well-being due to a smart wheelchair does not mean that this wheelchair should automatically be adopted. Well-being is only one value in the mix for adoption, where other values to consider would be human rights, respect, privacy, justice, freedom, culture, etc.

Computational Sustainability

Computational sustainability is an area of study within the A/IS community that demonstrates that the A/IS community is already showing interest in well-being even when not using this term, as the concept of sustainability encompasses aspects of well-being.

Computational sustainability directly relates to the use of these technologies to increase social good in ways that could be uniquely tied to existing well-being metrics. As defined by The Institute of Computational Sustainability, the field is designed to provide "computational models for a sustainable environment, economy, and society" and their project summary notes that:

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

Humanity's use of Earth's resources is threatening our planet and the livelihood of future generations. Computing and information science can — and should — play a key role in increasing the efficiency and effectiveness in the way we manage and allocate our natural resources. We propose an expedition in Computational Sustainability, encompassing computational and mathematical methods for a sustainable environment, economy, and society.

AAAI, (the Association for the Advancement of Artificial Intelligence) the world's largest global body dedicated to the advancement of artificial intelligence had a special track on computational sustainability at their 2017 conference. The description of the track provides helpful specifics demonstrating the direct alignment between the work of this Committee and the A/IS community at large:

> *This special track invites research papers on novel concepts, models, algorithms, and systems that address problems in computational sustainability. We are looking for a broad range of papers ranging from formal analysis to applied research. Examples include papers explaining how the research addresses specific computational problems, opportunities, or issues underlying sustainability challenges and papers describing a sustainability challenge or application that can be tackled using AI methods. Papers proposing general challenges and data sets for computational sustainability are also welcome. All AI topics that can address computational sustainability issues are appropriate, including machine learning, optimization, vision, and robotics, and others. Sustainability domains include natural resources, climate, and the environment (for example, climate change, atmosphere, water, oceans, forest, land, soil, biodiversity, species), economics and human behavior (for example, human well-being, poverty, infectious diseases, over-population, resource harvesting), energy (for example, renewable energy, smart grid, material discovery for fuel cell technology) and human-built systems (for example, transportation systems, cities, buildings, data centers, food systems, agriculture).*

Candidate Recommendations

- Work with influencers and decision-makers in the computational sustainability field to cross-pollinate efforts of computational sustainability in the A/IS field and the well-being communities to expedite efforts to identify, align, and advance robust and uniform indicators into current models that prioritize and increase human well-being. Develop cross-pollination between the computational sustainability and well-being professionals to ensure integration of well-being into computational sustainability, and vice-versa.
- Explore successful programs like LEED Building Design Standards, ISO 2600 Corporate Responsibility, ISO 37101 Sustainable Development Standards, and others to determine what new standards or certification models along these lines approach would be valuable and operationalizable for A/IS.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

Further Resources

- Gomes, C. P. "Computational Sustainability: Computational Methods for a Sustainable Environment, Economy, and Society" in *The Bridge: Linking Engineering and Society.* Washington, DC: National Academy of Engineering of the National Academies, 2009.
- Meadows, D. H., D. L. Meadows, J. Randers, and W. W. Behrens, III. The Limits to Growth. New York: Universe Books, 1972. Reissued in 2004 by Chelsea Green Publishing & Earthscan.
- LEED Building Design Standards program.
- ISO 2600, Guidance on Social Responsibility.
- ISO 37101, Sustainable Development in Communities

Issue:

Well-being indicators provide an opportunity for modeling scenarios and impacts that could improve the ability of A/IS to frame specific societal benefits for their use.

Background:

There is a lack of easily available or widely recognized scenarios along these lines.

Candidate Recommendation

Rigorously created well-being assessments could be utilized as a public "scoreboard," or statement of intent, that would provide innovation opportunities for technologists as well as a form of public accountability for human sustainability.

Further Resources

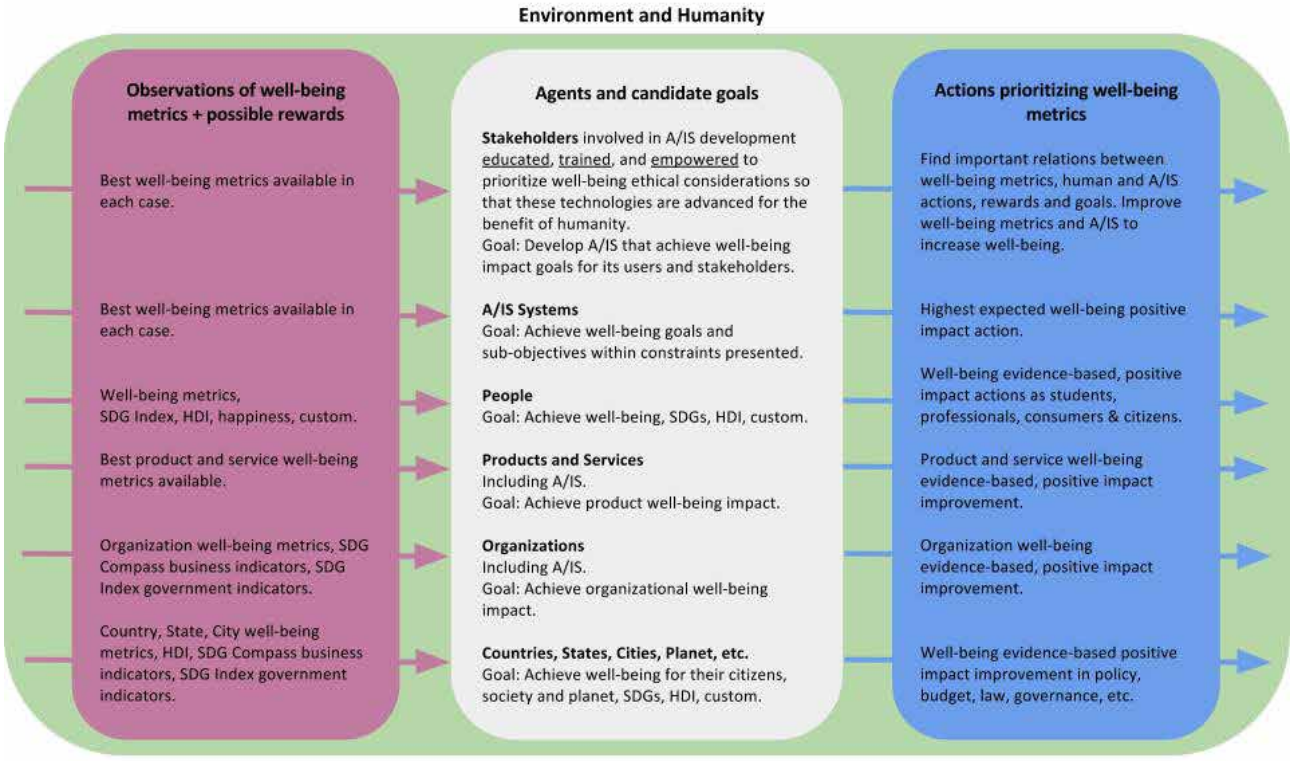
The following schema and well-being assessment tool provide an initial attempt to visualize how A/IS technologists can utilize well-being metrics in their work. By modeling the potential positive or negative impacts of technologies across a full spectrum of financial, environmental, and social impacts (e.g., a "triple bottom line" well-being indicator model) A/IS technologists can better avoid negative unintended consequences for human well-being, while increasing innovation and positive human well-being for their work.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

Schema of A/IS and the Stakeholders Involved in Their Development



Schema of A/IS systems and the stakeholders involved in their development, **adapting and operationalizing well-being metrics for ethical A/IS** in a world model.

The schema represents a model of the world where the stakeholders (designers, engineers, technologists, researchers, managers, users, etc.) involved in A/IS development adapt and operationalize well-being metrics for ethical A/IS. Stakeholders can visualize important entities in the world as agents with different goals that receive observations and possible rewards from the environment and make actions that could have positive and negative impacts to the well-being of different agents.

This schema could help to assess, in different cases, the well-being metrics that the A/IS should take into account and the well-being metrics of the impacts that A/IS actions could and can cause, related to important elements in the world like: people, products, organizations, climate, countries, etc. An applied case of this schema could be seen in the following well-being impact assessment.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

Well-being Impact Assessment

Here is a concept for simple A/IS well-being impact assessment, based on Maslow's Hierarchy of Need (where the Hierarchy would be considered an accredited and contextually appropriate metric of use). Given that a working definition of well-being including both individual and societal key performance indicators (KPIs) is still being developed, this metric is general and used for illustrative purposes only.

Please also note that this is a purely conceptual framework used as a directional teaching tool for readers. It doesn't yet include an evaluative component or reflect the holistic nature of well-being at this time like The Happiness Screening Tool (based on the government of Bhutan's Policy Screening Tool) provided in the Appendix. It should be noted that any impact assessment created by A/IS and well-being experts working together identify best-in-class (existing) metrics within specific contexts of use.

	Individual Direct	Individual Indirect	Environment Direct	Individual Indirect	Social Direct	Social Indirect
Basic Needs						
Safety						
Belonging						
Esteem						
Self-Actualization						
Overall Impact						

Indicators:

nil impact = 0 *negative impact = –* *positive impact = +* *unknown impact = ?*

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

The following examples are provided to demonstrate specific A/IS applications within this framework and include: a retail kiosk robot, a small factory arm, a mental health chatbot, and a companion robot. The goal of these diagrams is to provide a sample of how the work of matching established well-being metrics to A/IS work could progress.

Retail Kiosk Robot	Individual Direct	Individual Indirect	Environment Direct	Individual Indirect	Social Direct	Social Indirect
Basic Needs	0	0	0	–	+	?
Safety	?	?	?	–	+	?
Belonging	+	?	+	?	+	?
Esteem	+	?	0	0	+	?
Self-Actualization	?	?	0	0	?	?
Overall Impact	Mild +	Unknown	Very Mild +	Mild –	Strong +	Unknown

Using this tool, the retail kiosk robot scores are mildly beneficial in the category of Individual Direct (i.e., reduced barriers to goal attainment) and Environmental Direct (i.e., use of resources), while strongly beneficial in Social Direct (i.e., better access to mental health support), but mildly unbeneficial in Environment Indirect (i.e., carbon footprint), and unknown in Social Indirect (i.e., job loss) categories. The robot is "helpful and kind," but of limited utility or interaction value. Another example of a negative impact on well-being is gendering, racial identification, or physical attributes of kiosk robots (such as a slim, youthful appearing, Caucasian, female), leading to harmful stereotyping.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

Small Factory Arm	Individual Direct	Individual Indirect	Environment Direct	Individual Indirect	Social Direct	Social Indirect
Basic Needs	+	+	0	–	+	+
Safety	?	?	?	–	?	+
Belonging	–	–	0	0	0	0
Esteem	–	–	0	0	0	0
Self-Actualization	0	0	0	0	0	0
Overall Impact	Mild –	Mild –	Nil	Mild –	Mild +	Mild +

The tool indicates that robots need to be assessed more thoroughly on their safe operations to better answer impact assessment, and that this is also a robot with very limited interaction with people. But the diagram shows how the arm could have a potentially negative impact on self-worth and belonging, but a positive impact on basic needs both for individuals and society.

Mental Health Chatbot	Individual Direct	Individual Indirect	Environment Direct	Individual Indirect	Social Direct	Social Indirect
Basic Needs	0	0	0	0	0	0
Safety	+	0	0	0	?	+
Belonging	+	?	0	0	?	–
Esteem	+	?	0	0	?	–
Self-Actualization	?	0	0	0	0	0
Overall Impact	Strong +	Unknown	Nil	Nil	Unknown	Mild –

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

There is evidence that a mental health aide chatbot could improve individual self esteem and ultimately reduce self harm, but there is little evidence supporting claims that this would improve society directly or indirectly. The reliance on artificial support may have a net negative impact on society. However, this would need to be determined by the A/IS and well-being experts applying this methodology once created in a robust and rigorous manner.

Companion Robot like Paro	Individual Direct	Individual Indirect	Environment Direct	Individual Indirect	Social Direct	Social Indirect
Basic Needs	0	0	0	-	0	0
Safety	+	?	0	0	0	0
Belonging	+	?	0	0	?	-
Esteem	+	?	0	0	?	-
Self-Actualization	?	0	0	0	0	0
Overall Impact						

For a small resource cost, a companion robot can provide significant psychological assistance. On the one hand, this makes society more caring, but on the other hand reliance on artificial companionship shows a lack of social resources in this area. A potential negative impact is development of reliance on companionship and negative impact on people who lose access to companion robot.

The Happiness Project Screening Tool for Business provided in the Appendix could also augment this if a product shows a low or negative score in the areas of well-being. Another set of metrics that could be used in a more detailed schema are the Kingdom of Bhutan's nine domains of well-being: psychological well-being, health, community vitality, living standards, governance, environment diversity, culture, education, and time use.

Whatever established well-being metrics that may be utilized for such a methodology, it is critical for A/IS technologists and well-being experts to work in unison to create assessment tools using best in class data, indicators, and practices in their potential analysis and use.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Section 3 — Adaptation of Well-being Metrics for A/IS

This section focuses on areas of immediate attention for A/IS technologists to be aware of regarding well-being metrics in an effort to aid their work and avoid negative unintended consequences.

Issue:

How can creators of A/IS incorporate measures of well-being into their systems?

Background

Just as undirected A/IS can lead to negative outcomes, A/IS directed only to specific ends without considering human well-being can lead to negative side effects. Without practical ways of incorporating widely shared ways of measuring and promoting well-being metrics and expected well-being outcomes available to designers, A/IS will likely lack beneficence.

Once well-being metrics are widely recognized as a directional requirement for society, conceptually, one would like such measures to be supported by the engines of change and leverage within society. A/IS will be an increasing portion of such engines. How might designers architect systems to include such measures as considerations while executing their primary objectives? How will these measures be adapted as we learn more?

Existing metrics of well-being could be formulated into a sub-objective of the A/IS. In order to operate with respect to such sub-objectives, it is instrumental to evaluate the consequences of the A/IS's actions. As practical systems are bounded and can predict over only limited horizons, it may be necessary to supplement these evaluations with both biases toward virtues and deontological guidelines or soft constraints as lesser supplemental components, informed by the well-being metrics and their precursors or constituents.

As these well-being sub-objectives will be only a subset of the intended goals of the system, the architecture will need to balance multiple objectives. Each of these sub-objectives may be expressed as a goal, or as a set of rules, or as a set of values, or as a set of preferences, and those can be combined as well, using established methodologies from intelligent systems engineering.

For example, people, organizations, and A/IS, collaborating together, could understand the well-being impacts and objectives of products, services, organizations, and A/IS within the context of the well-being of communities, cities, countries, and the planet using the SDG Index

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

and Dashboards, the SDG Compass Inventory of Business Indicators and other metrics. This collaboration of people, organizations and A/IS could make decisions and take actions with high expected utility to well-being objectives and goals such as those stated in the Sustainable Development Goals and similar institutions. This collaboration could lead to a more humane, organizational, and computational sustainability for individuals, all of society, and the planet.

International organizations, lawmakers, and policy experts can specify core values and/or sub-objectives as rules for the benefit of society utilizing well-being metrics as a starting point and these can be pluggable and hierarchical by jurisdiction. Similarly, industry organizations would be able to specialize norms and industry self-regulation (e.g., any automated flight attendants should prevent onboard smoking and sit down during takeoff) as a layer.

System designers should ensure situational awareness as well as prediction of the consequences of their actions based on some world model. They could also layer in their own sub-objectives and make the system's values explicit.

Resellers, service organizations, or owners that have particular primary goals for their systems would still be able to specify primary goals for the system (e.g., mowing lawns, doing taxes, etc.), and those would be alongside the other deeper-down subgoals and values as well for societal benefit, public safety, etc., directly relating to established well-being metrics.

End users would have the opportunity to layer on their own preferences in these systems, and would also be able to get an explanation and inventory of the types of objectives or value systems the A/IS holds relating to established well-being metrics, including what permissioning is required for modifying or removing them.

Candidate Recommendation

Formation of a working group to develop a blueprint for the fluid and evolving (institutional learning) operationalization of A/IS well-being indicators for the various stakeholders (e.g., technicians, coders, and system designers), international well-being oriented organizations, lawmakers, and policy experts, industry organizations, retailers, resellers, service organizations and owners, and end users.

Candidate Recommendation

Creation of technical standards for representing dimensions, metrics, and evaluation guidelines for well-being metrics and their precursors and constituents within A/IS. This would include ontologies for representing requirements as well as a testing framework for validating adherence to well-being metrics and ethical principles. (For more information, please see IEEE P7010™ Standards Working Group mentioned above).

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

Further Resources

- Calvo, R. A., and D. Peters. Positive Computing: Technology for Well-Being and Human Potential. Cambridge MA: MIT Press, 2014
- Collette Y., and P. Slarry. Multiobjective Optimization: Principles and Case Studies (Decision Engineering Series). Berlin, Germany: Springer, 2004. doi: 10.1007/978-3-662-08883-8.
- Greene, J. et al. "Embedding Ethical Principles in Collective Decision Support Systems," in: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence,* 4147–4151. Palo Alto, CA: AAAI Press, 2016.
- Li, L. et al. "An Ontology of Preference-Based Multiobjective Evolutionary Algorithms," 2016. CoRR abs/1609.08082.
- A. FT Winfield, C. Blum, and W. Liu. "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection," in *Advances in Autonomous Robotics Systems*. Springer, 2014, pp. 85–96.
- Gershman, S. J., E. J. Horvitz, and J. B. Tenenbaum. "Computational rationality: A converging paradigm for intelligence in brains, minds, and machines." *Science* 349, no. 6245 (2015): 273–278.
- PositiveSocialImpact: Empowering people, organizations and planet with information and knowledge to make a positive impact to sustainable development.

Issue:

A/IS technologies designed to replicate human tasks, behavior, or emotion have the potential to either increase or decrease well-being.

Background

A/IS are already being executed in ways that could dramatically increase human well-being or, possibly, have an undue coercive effect on humans.

A/IS technologies present great opportunity for positive change in every aspect of society. However, sophisticated manipulative technologies utilizing A/IS can also restrict the fundamental freedom of human choice, and are able to manipulate humans who consume customized content without recognizing the extent of manipulation. Software platforms are moving from targeting content to much more powerful and potentially harmful "persuasive computing." A/IS with sophisticated manipulation technologies (so-called "big nudging") will be able to guide individuals through entire courses of action, whether it be a complex work process, consumption of free content, or political persuasion.

There is also a related concern that big nudging can be done without anyone realizing harm is occurring. With deep learning methods,

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

technologies may not be understood, much less contemplated. This begs the age-old question: just because one can do something, does that mean one should? Hence, there is a need to understand A/IS well-being related processes and impacts further, and to devise ways to protect people from harm and secure well-being in the furtherance of A/IS.

A/IS may also deceive and harm humans by posing as humans. With the increased ability of artificial systems to meet the Turing test (an intelligence test for a computer that allows a human to distinguish human from artificial intelligence), there is a significant risk that unscrupulous operators will abuse the technology for unethical commercial, or outright criminal, purposes. The widespread manipulation of humans by A/IS and loss of human free agency, autonomy, and other aspects of human flourishing, is by definition a reduction in human well-being. Without taking action to prevent it, it is highly conceivable that A/IS will be used to deceive humans by pretending to be another human being in a plethora of situations or via multiple mediums.

Without laws preventing A/IS from simulating humans for purposes like deception and coercion, and enforcing A/IS to clearly identify as such, mistaken identity could also reasonably be expected.

Candidate Recommendation

To avoid potential negative unintended consequences, A/IS manufacturers, and society in general, should prioritize the analysis and implementation of practices and policy that secures or increases human well-being, including:

- Well-being metrics to guide the development and implementation of A/IS should increase human well-being, defined subjectively in terms of cognitive, affective, and eudaimonic domains, and objectively in terms of conditions enabling well-being.
- While individuals may enjoy the ability of A/IS to simulate humans in situations where they are pure entertainment, explicit permission and consent by users in the use of these systems is recommended, and the well-being impacts on users should be monitored, researched, and considered by the A/IS community in an effort to provide services and goods that improve well-being. As part of this, it is important to include multiple stakeholders, including minorities, the marginalized, and those often without power or a voice.
- The implications of A/IS on human well-being are important issues to research and understand. A literature review to determine the status of academic research on the issue of A/IS impacts on human well-being needs to be conducted and aggregated in a centralized repository for the A/IS community.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

Further Resources

- Helbing, D. et al. "Will Democracy Survive Big Data and Artificial Intelligence?" *Scientific American*, February 25, 2017.
- Schenker, J. L. "Can We Balance Human Ethics with Artificial Intelligence?" *Techonomy,* January 23, 2017.
- Bulman, M. "EU to Vote on Declaring Robots To Be 'Electronic Persons." *Independent,* January 14, 2017.
- Nevejan, N. for the European Parliament. "European Civil Law Rules in Robotics." October 2016.
- "The AI That Pretends To Be Human," *LessWrong* blog post, February 2, 2016.
- Chan, C. "Monkeys Grieve When Their Robot Friend Dies." *Gizmodo*, January 11, 2017.

Issue:

Human rights law is sometimes conflated with human well-being, leading to a concern that a focus on human well-being will lead to a situation that minimizes the protection of inalienable human rights, or lowers the standard of existing legal human rights guidelines for non-state actors.

Background

International human rights law has been firmly established for decades and the protection of human rights must be an end result in itself. Some countries or regimes have highlighted the use or increase of certain "well-being" measures as justification to violate human rights, as happens in countries that conduct ethnic cleansing or mistreat refugees or immigrants who are portrayed as threatening a nation's culture or economic structure.

While the use of well-being metrics to justify human rights violations is an unconscionable perversion of the nature of any well-being metric, these same practices happen today in relation to the GDP. For instance, today, according to the International Labor Organization (ILO) approximately 21 million people are victims of forced labor (slavery) representing between 9% to 56% of various countries current GDP income. These clear human rights violations, from sex trafficking and child armies, to indentured farming or manufacturing labor, increase a country's GDP.

Well-being metrics and mechanisms should also take into consideration, and happen in conjunction with, independent assessments on respect and international obligations to promote, protect, and fulfill a full spectrum of human rights. For example, the use of the goal of well-being in the context of repairing and enhancing humans, predictive policing, or autonomous weapons systems to protect the public may have negative impacts on the rights of individuals or groups. Moreover, the development and delivery of A/IS should adopt a human rights approach to technology, including, but not limited to, the

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

UN Guiding Principles on Human Rights (also known as the Ruggie principles).

To avoid issues of conflation and confusion, it is critical to note the following: human rights involves adhering to the firmly established application of international human rights law. Well-being metrics are designed to measure the efficacy of the implementation of methodologies and policy related to individual and societal flourishing.

Well-being as a value is also distinct from justice, responsibility, and freedom. But A/IS technologies can be narrowly conceived from an ethical standpoint and still be legal and safe in their usage following existing practices, but not contribute to human well-being. In this regard, well-being considerations do not displace other issues of human rights or ethical methodologies, but rather complement them.

Candidate Recommendation

Human rights and human well-being should not be held as trade-offs, with one to be prioritized over the other. In this regard, well-being metrics can be complementary to the goals of human rights, but cannot and should not be used as a proxy for human rights or any existing law.

Further Resources

- Project Include - The site features an open source manual for creating diversity in tech and highlights three key points for creating change: inclusion, comprehensiveness, and accountability.
- OpenDiversityOrg initiative from Double Union and Project Include have an action document with a lot of recommendations.
- "The Diversity Debt" by Susan Wu at Project Include is a compelling example of converting a problem into innovation language.

Issue:

A/IS represents opportunities for stewardship and restoration of natural systems and securing access to nature for humans, but could be used instead to distract attention and divert innovation until the planetary ecological condition is beyond repair.

Background

Human well-being, the existence of many other species, as well as economic and social systems, draw from and depend upon healthy ecological systems and a healthy local and planetary environment. Research using geo-data finds that human well-being is enhanced through access to nature. Many bank on technology to answer the threats of climate change, water scarcity, soil degradation, species extinction, deforestation, deterioration of biodiversity, and destruction of ecosystems that threaten humankind and other life forms.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

While technology may be the answer for some of these threats, it is unclear whether benefits extend beyond those from high socio-economic class to the majority of people, particularly the middle class and working poor, as well as those suffering from abject poverty, fleeing disaster zones or otherwise lacking the resources to meet their needs. For example, in cities in China where air pollution is so prevalent that the air is unhealthy, a few schools have covered "outdoor" fields with domes full of purified air while most children must risk their lungs when playing outside, or play indoors. Moreover, it is well-understood that ecological crises, such as sea level rise and fisheries depletion, will not only negatively impact business interests, but it will have a significantly more devastating impact on the poor and developing nations than the wealthy and developed nations.

Candidate Recommendation

Well-being metrics employed for A/IS should include measures for ecological/environmental sustainability that point the direction toward stewardship and restoration of natural systems and ensure equitable environmental justice.

Candidate Recommendation

Convene a committee to issue findings on the modalities and potentials already identified in which A/IS makes progress toward stewardship and restoration of natural systems; trends in the A/IS field that represent threats to and opportunities for ecological sustainability and environmental justice; and areas for suggested future innovation and implementation.

Further Resources

- Newton, J. "Well-being and the Natural Environment: An Overview of the Evidence." August 20, 2007.
- Dasgupta, P. Human Well-Being and the Natural Environment. Oxford, U.K.: Oxford University Press, 2001.
- Haines-Young, R., and M. Potschin. "The Links Between Biodiversity, Ecosystem Services and Human Well-Being," in *Ecosystem Ecology: A New Synthesis*, edited by D. Raffaelli, and C. Frid. Cambridge, U.K.: Cambridge University Press, 2010.
- Hart, S. *Capitalism at the Crossroads: Next Generation Business Strategies for a Post-Crisis World.* Upper Saddle River, NJ: Pearson Education, 2010.
- United Nations Department of Economic and Social Affairs. "Call for New Technologies to Avoid Ecological Destruction." Geneva, Switzerland, July 5, 2011.
- Pope Francis. Encyclical Letter Laudato Si, On the Care of Our Common Home. May 24, 2015.

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

Issue:

The well-being impacts of A/IS applied to human genomes are not well understood.

Background

As A/IS are increasingly used to interpret the health significance of our genomics data ("deep genomics") and to contribute to the subsequent engineering and editing of our genomes, important ethical and governance questions are in the background that provide an opportunity to utilize well-being metrics to ensure the beneficial development of genomic research as it relates to A/IS.

Imagine this scenario:

6 A.M., Washington, DC — Erika wakes up and quickly checks her "digital DNA avatar," a digital version of her genetic blueprint as it evolves day by day. The avatar knows a lot about her as it constantly monitors the interactions between her genes, analyzes her bodily fluids and diet, as well as integrates data about the air quality around her. Her avatar proposes a few advices about food choices and exercise patterns. Everything seems in check, nothing to be worried about. For now.

A first overarching reflection concerns the relationship between well-being and an increasing ability to understand and engineer our genomes: How do in-depth and personalized understanding of how our genomes function and evolve relate to the notion of well-being as measured traditionally and/or according to well-being measures? When does a reductionist interpretation of the health significance of our genomics data threaten our well-being?

Other significant questions include:

- How accurate will the predictive health data coming from the convergence of A/IS and genomics be?
- How will these health predictions be used, and who will have access to them?
- Do pharmaceutical and insurance companies have the right to use and profit from your health data predictions/modeling without giving you any benefits back in return?
- Would it threaten your self-worth if those handling your health data know a lot of biological details about your body?
- Is it ethical for a prospective employer to ask how your health will look like in the next decade?

Answers to these questions are not easy to capture, but their impact on well-being within society is profound.

The convergence of genomics technologies and A/IS creates new opportunities to define our identity and well-being within a simple narrative in which our genes have the power to tell us who

creative commons

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.

Well-being

and how well we are. As A/IS are increasingly used to interpret the health significance of our genomics data ("deep genomics") and to contribute to the subsequent engineering/editing of our genomes, we should consider important ethical and governance questions.

There is an urgent need to concurrently discuss how the convergence of A/IS and genomic data interpretation will challenge the purpose and content of relevant legislation that preserve well-being, such as, for the United States, the Health Insurance Portability and Accountability Act (HIPAA) and the Genetic Information Non-Discrimination Act (GINA). Finding the right balance of protection and regulation in using A/IS to interpret the health significance of genomics data will be important. Too much regulation could endanger precision medicine initiatives in some countries, while others would be leading the bio-race. Too little regulation could leave citizens vulnerable to different forms of threats to their well-being.

Candidate Recommendation

A working committee should be convened gathering those at the sharp end of genomics, A/IS, ethics, and governance to start a conversation with different communities to better understand the impact on well-being of the use of A/IS to interpret (and engineer) genomics data.

Candidate Recommendation

Relevant expert and legislative committees should commission a study on the impact on well-being of deep genomics, meaning at the convergence of genomics and A/IS. Such a study is recommended to encompass diverse fields of expertise in philosophy, sociology, ethics, biosafety, biosecurity, and genomics governance. Recommendations from the study should draft proposals to frame debates in legislatures and help lawmakers start developing appropriate legislation to govern A/IS applied to genomes for the well-being of society.

creative commons
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 United States License.