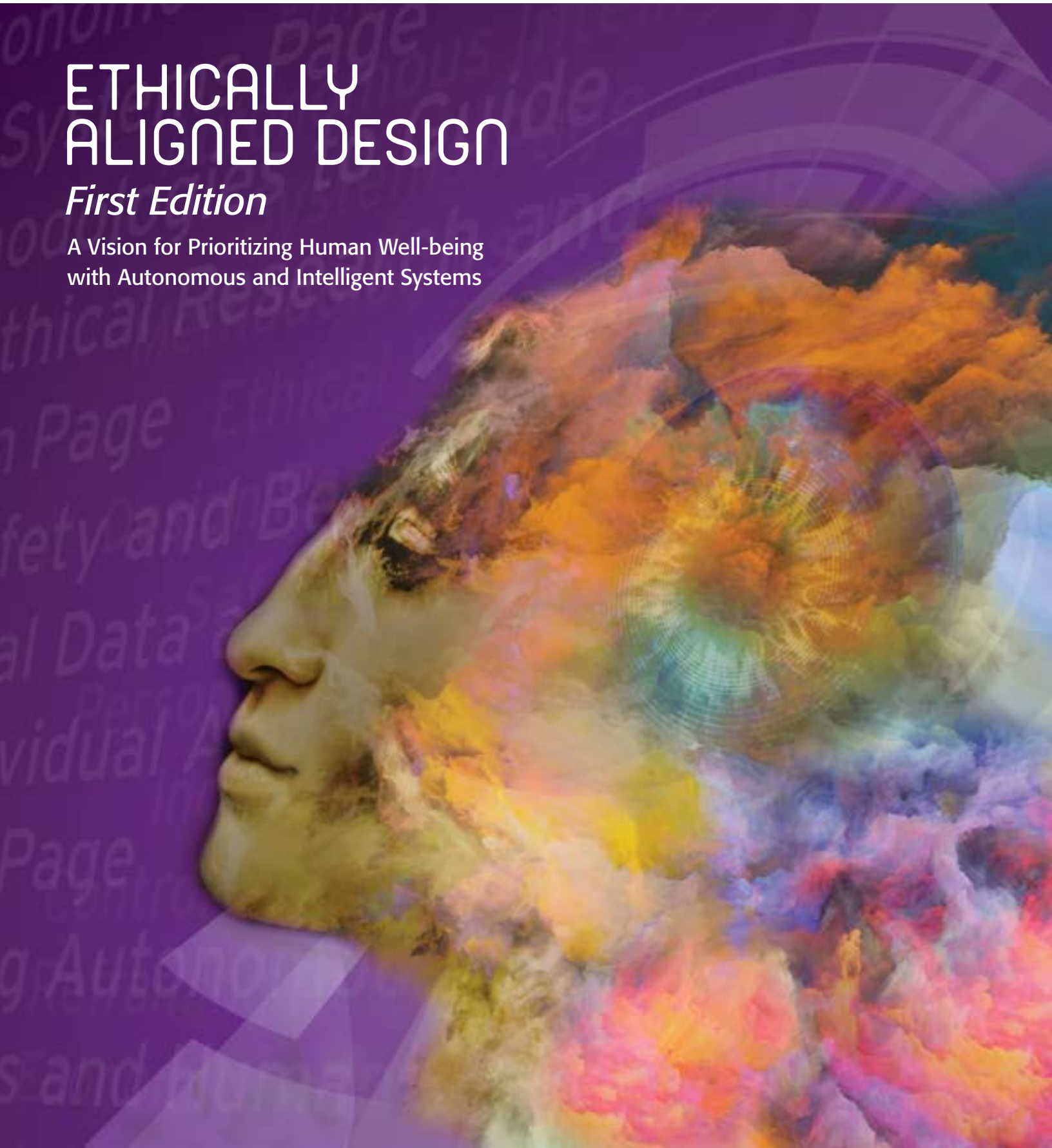
IEEE

Advancing Technology for Humanity

ETHICALLY ALIGNED DESIGN

First Edition

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



The views and opinions expressed in this collaborative work are those of the authors and do not necessarily reflect the official policy or position of their respective institutions or of the Institute of Electrical and Electronics Engineers (IEEE). This work is published under the auspices of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems for the purposes of furthering public understanding of the importance of addressing ethical considerations in the design of autonomous and intelligent systems.

Please see page 290, How the Document Was Prepared, for more details regarding the preparation of this document.

Table of Contents

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 must be developed and should operate in a way that is beneficial to people and the environment, beyond simply reaching functional goals and addressing technical problems. This approach will foster the heightened level of trust between people and technology that is needed for its fruitful 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 explicit or implicit values, including our imaginary[1] around so-called "Artificial Intelligence" and the institutions, symbols, and representations it generates.

Ultimately, our goal should be *eudaimonia*, a practice elucidated by Aristotle that defines human well-being, both at the individual and collective level, as the highest virtue for a society. Translated roughly as "flourishing", the benefits of eudaimonia begin with conscious contemplation, where ethical considerations help us define how we wish to live.

Whether our ethical practices are Western (e.g., Aristotelian, Kantian), Eastern (e.g., Shinto, 墨家/School of Mo, Confucian), African (e.g., Ubuntu), or from another tradition, honoring holistic definitions of societal prosperity is essential versus pursuing one-dimensional goals of increased productivity or gross domestic product (GDP). Autonomous and intelligent systems should prioritize and have as their goal the explicit honoring of our inalienable fundamental rights and dignity as well as the increase of human flourishing and environmental sustainability.

The goal of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems ("The IEEE Global Initiative") is that *Ethically Aligned Design* will provide pragmatic and directional insights and recommendations, serving as a key reference for the work of technologists, educators and policymakers in the coming years.

Ethically Aligned Design sets forth scientific analysis and resources, high-level principles, and actionable recommendations. It offers specific guidance for standards, certification, regulation or legislation for design, manufacture, and use of A/IS that provably aligns with and improves holistic societal well-being.

[1]The symbols, values, institutions, and norms of a societal group through which people imagine their lives and constitute their societies.

Executive Summary

I. Purpose of *Ethically Aligned Design, First Edition* (EAD1e)

Autonomous and intelligent technical systems are specifically designed to reduce the necessity for human intervention in our day-to-day lives. In so doing, these new systems are also raising concerns about their impact on individuals and societies. Current discussions include advocacy for a positive impact, such as optimization of processes and resource usage, more informed planning and decisions, and recognition of useful patterns in big data. Discussions also include warnings about potential harm to privacy, discrimination, loss of skills, adverse economic impacts, risks to security of critical infrastructure, and possible negative long-term effects on societal well-being.

Because of their nature, the full benefit of these technologies will be attained only if they are aligned with society's defined values and ethical principles. Through this work we intend, therefore, to establish frameworks to guide and inform dialogue and debate around the non-technical implications of these technologies, in particular related to ethical aspects. We understand "ethical" to go beyond moral constructs and include social fairness, environmental sustainability, and our desire for self-determination.

Our analyses and recommendations in *Ethically Aligned Design* address values and intentions as well as implementations, both legal and technical. They are both aspirational, what we hope or wish should happen, and practical, what we—the techno-scientific community and every group involved with and/or affected by these technologies—could do for society to advance in positive directions. The analyses and recommendations in EAD1e are offered as guidance for consideration by governments, businesses, and the public at large in the advancement of technology for the benefit of humanity.

Chapters in *Ethically Aligned Design*, First Edition

1. From Principles to Practice
2. General Principles
3. Classical Ethics in A/IS
4. Well-being
5. Affective Computing
6. Personal Data and Individual Agency
7. Methods to Guide Ethical Research and Design
8. A/IS for Sustainable Development
9. Embedding Values into Autonomous and Intelligent Systems
10. Policy
11. Law

Introduction

II. General Principles

The ethical and values-based design, development, and implementation of autonomous and intelligent systems should be guided by the following General Principles:

1. **Human Rights**
   A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
2. **Well-being**
   A/IS creators shall adopt increased human well-being as a primary success criterion for development.
3. **Data Agency**
   A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
4. **Effectiveness**
   A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
5. **Transparency**
   The basis of a particular A/IS decision should always be discoverable.
6. **Accountability**
   A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
7. **Awareness of Misuse**
   A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
8. **Competence**
   A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

III. Ethical Foundations

Classical Ethics

By drawing from over two thousand five hundred years of classical ethics traditions, the authors of *Ethically Aligned Design* explored 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, this work 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.

IV. Areas of Impact

A/IS for Sustainable Development

Through affordable and universal access to communications networks and the Internet, autonomous and intelligent 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 address humanitarian and sustainable development issues resulting in increased individual societal and environmental well-being. Such efforts could be facilitated through the recognition of and adherence to established indicators of societal flourishing such as the United Nations Sustainable Development Goals so that human well-being is utilized as a primary success criteria for A/IS development.

Introduction

Personal Data Rights and Agency Over Digital Identity

People have the right to access, share, and benefit from their data and the insights it provides. Individuals require mechanisms to help create and curate the terms and conditions regarding access to their identity and personal data, and to control its safe, specific, and finite exchange. Individuals also require policies and practices that make them explicitly aware of consequences resulting from the aggregation or resale of their personal information.

Legal Frameworks for Accountability

The convergence of autonomous and 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 even a human physical appearance. The issue of the legal status of complex autonomous and intelligent systems thus intertwines with broader legal questions regarding how to ensure accountability and allocate liability when such systems cause harm. It is clear that:

- Autonomous and intelligent 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. These stakeholders should adopt rules and standards that ensure effective human control over those decisions and how to allocate legal responsibility for harm caused by them.
- The manifestations generated by autonomous and intelligent technical systems should, in general, be protected under national and international laws.
- Standards of transparency, competence, accountability, and evidence of effectiveness should govern the development of autonomous and intelligent systems.

Policies for Education and Awareness

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

- Support, promote, and enable internationally recognized legal norms.
- Develop government expertise in related technologies.
- Ensure governance and ethics are core components in research, development, acquisition, and use.
- Regulate to ensure public safety and responsible system design.
- Educate the public on societal impacts of related technologies.

Introduction

V. Implementation

Well-being Metrics

For autonomous and intelligent systems to provably advance a specific benefit for humanity, there need to be clear indicators of that benefit. Common metrics of success include profit, gross domestic product, consumption levels, and occupational safety. While important, these metrics fail to encompass the full spectrum of well-being for individuals, the environment, and society. Psychological, social, economic fairness, and environmental factors matter. Well-being metrics include 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. A/IS can improve capturing of and analyzing the pertinent data, which in turn could help identify where these systems would increase human well-being, providing new routes to societal and technological innovation.

Embedding Values into Autonomous and Intelligent Systems

If machines engage in human communities as quasi-autonomous agents, then those agents must be expected to follow the community's social and moral norms. Embedding norms in such quasi-autonomous 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.

Methods to Guide Ethical Research and Design

To create autonomous and intelligent technical systems that enhance and extend human well-being and freedom, values-based design methods must put human advancement at the core of development of technical systems. This must be done in concert with the recognition that machines should serve humans and not the other way around. Systems developers should employ values-based design methods in order to create sustainable systems that can be evaluated in terms of not only providing increased economic value for organizations but also of broader social costs and benefits.

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 lives. To ensure that intelligent technical systems will be used to help humanity to the greatest extent possible in all contexts, autonomous and intelligent systems that participate in or facilitate human society should not cause harm by either amplifying or dampening human emotional experience.

Acknowledgements

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

We wish to thank the Executive Committee and Committees of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems:

Executive Committee Officers

Raja Chatila, *Chair*
Kay Firth-Butterfield, *Vice Chair*
John C. Havens, *Executive Director*

Executive Committee Members

Dr. Greg Adamson, Karen Bartleson, Virginia Dignum, Danit Gal, Malavika Jayaram, Sven Koenig, Eileen M. Lach, Raj Madhavan, Richard Mallah, AJung Moon, Monique Morrow, Francesca Rossi, Alan Winfield, and Hagit Messer Yaron

Committee Chairs

- **General Principles:** Mark Halverson, and Peet van Biljon
- **Embedding Values into Autonomous Intelligent Systems:** Francesca Rossi and Bertram F. Malle
- **Methodologies to Guide Ethical Research and Design:** Raja Chatila and Corinne Cath
- **Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI):** Malo Bourgon and Richard Mallah
- **Personal Data and Individual Agency:** Katryna Dow and John C. Havens
- **Reframing Autonomous Weapons Systems:** Peter Asaro
- **Sustainable Development:** Elizabeth Gibbons
- **Law:** Nicolas Economou and John Casey
- **Affective Computing:** John Sullins and Joanna J. Bryson
- **Classical Ethics in A/IS:** Jared Bielby
- **Policy:** Peter Brooks and Mina Hannah
- **Extended Reality:** Monique Morrow and Jay Iorio
- **Well-being:** Laura Musikanski and John C. Havens
- **Editing:** Karen Bartleson and Eileen M. Lach
- **Outreach:** Maya Zuckerman and Ali Muzaffar
- **Communications:** Leanne Seeto and Mark Halverson
- **High School:** Tess Posner
- **Global Coordination:** Victoria Wang, Arisa Ema, Pavel Gotovtsev

Introduction

Programs and Projects Inspired by The IEEE Global Initiative:

- ***Ethically Aligned Design* University Consortium:** Hagit Messer, *Chair*
- ***Ethically Aligned Design* Community:** Lisa Morgan, Program Director, Content and Community
- **Ethics Certification Program for Autonomous and Intelligent Systems:** Meeri Haataja, Chair; Ali Hessami, *Vice-Chair*
- **Glossary:** Sara M. Jordan, *Chair*

People

We would like to warmly recognize the leadership and constant support of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems by Dr. Ing. Konstantinos Karachalios, Managing Director of the IEEE Standards Association.

We would also like to thank Stephen Welby, Executive Director and Chief Operating Officer of IEEE for his generous and insightful support of the *Ethically Aligned Design*, First Edition process and The IEEE Global Initiative overall.

We would especially like to thank Eileen M. Lach, the former IEEE General Counsel and Chief Compliance Officer, whose 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 served as a strong catalyst for the development of the Initiative within IEEE.

Finally, we would like to also acknowledge the ongoing work of three Committees of The IEEE Global Initiative regarding their chapters of *Ethically Aligned Design* that, for timing reasons, we were not able to include in *Ethically Aligned Design*, First Edition. These Committees include: Reframing Autonomous Weapons Systems, Extended Reality (formerly Mixed Reality) and Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI). We would like to thank Peter Asaro, Monique Morrow and Jay Iorio, Malo Bourgon and Richard Mallah for their leadership in these groups along with all their Committee Members. Once these chapters have completed their review and been accepted by IEEE they could either be included in *Ethically Aligned Design*, published by The IEEE Global Initiative, or in other publications of IEEE.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

From Principles to Practice
Ethically Aligned Design Conceptual Framework

Ethically Aligned Design, First Edition (EAD1e) represents more than a comprehensive report, distilling the consensus of its vast community of creators into a set of high-level ethical principles, key issues, and practical recommendations. EAD1e is an in-depth seminal work, a one-of-a-kind treatise, intended not only to inform a broader public but also to inspire its audience and readership of academics, engineers, policy makers, and manufacturers of autonomous and intelligent systems[1] (A/IS) to take action.

This Chapter, "From Principles to Practice", provides a mapping of the conceptual framework of *Ethically Aligned Design*. It outlines the logic behind "Three Pillars" that form the basis of EAD1e, and it connects the Pillars to high-level "General Principles" which guide all manner of ethical A/IS design. Following this, the content of the Chapters of EAD1e is mapped to the Principles. Finally, examples of EAD1e already in practice are described.

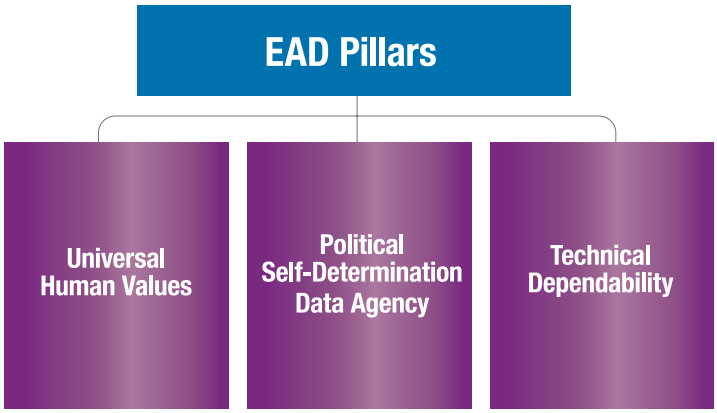
Sections in this Chapter:

- The Three Pillars of the *Ethically Aligned Design* Conceptual Framework
- The General Principles of *Ethically Aligned Design*
- Mapping the Pillars to the Principles
- Mapping the Principles to the Content of the Chapters
- From Principles to Practice
- *Ethically Aligned Design* in Implementation

The Three Pillars of the *Ethically Aligned Design* Conceptual Framework

The Pillars of the *Ethically Aligned Design* Conceptual Framework fall broadly into three areas, reflecting anthropological, political, and technical aspects:

1. **Universal Human Values:** A/IS can be an enormous force for good in society provided they are designed to respect human rights, align with human values, and holistically increase well-being while empowering as many people as possible. They should also be designed to safeguard our environment and natural resources. These values should guide policy makers as well as engineers, designers, and developers. Advances in A/IS should be in the service of all people, rather than benefiting solely small groups, a single nation, or a corporation.
2. **Political Self-Determination and Data Agency:** A/IS—if designed and implemented properly—have a great potential to nurture political freedom and democracy, in accordance with the cultural precepts of individual societies, when people have access to and control over the data constituting and representing their identity. These systems can improve government effectiveness and accountability, foster trust, and protect our private sphere, but only when people have agency over their digital identity and their data is provably protected.
3. **Technical Dependability:** Ultimately, A/IS should deliver services that can be trusted.[2] This trust means that A/IS will reliably, safely, and actively accomplish the objectives for which they were designed while advancing the human-driven values they were intended to reflect. Technologies should be monitored to ensure that their operation meets predetermined ethical objectives aligning with human values and respecting codified rights. In addition, validation and verification processes, including aspects of explainability, should be developed that could lead to better auditability and to certification[3] of A/IS.



The General Principles of *Ethically Aligned Design*

The General Principles of *Ethically Aligned Design* have emerged through the continuous work of dedicated, open communities in a multi-year, creative, consensus-building process. They articulate high-level principles that should apply to all types of autonomous and intelligent systems (A/IS). Created to guide behavior and inform standards and policy making, the General Principles define imperatives for the ethical design, development, deployment, adoption, and decommissioning of autonomous and intelligent systems. The Principles consider the role of A/IS creators, i.e., those who design and manufacture, of operators, i.e., those with expertise specific to use of A/IS, other users, and any other stakeholders or affected parties.

The General Principles[4] of *Ethically Aligned Design*

1. **Human Rights–**A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
2. **Well-being–**A/IS creators shall adopt increased human well-being as a primary success criterion for development.
3. **Data Agency–**A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
4. **Effectiveness–**A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
5. **Transparency–**The basis of a particular A/IS decision should always be discoverable.
6. **Accountability–**A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
7. **Awareness of Misuse–**A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
8. **Competence–**A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

Mapping the Pillars to the Principles

Whereas the Pillars of the *Ethically Aligned Design* Conceptual Framework represent broad anthropological, political, and technical aspects relating to autonomous and intelligent systems, the General Principles provide contextual filters for deeper analysis and pragmatic implementation.

It is also important to recognize that the General Principles do not live in isolation of EAD's Pillars and vice versa. While the General Principle of "Transparency" may inform the design of a specific autonomous or intelligent system, the A/IS must also account for universal human values, political self-determination, and data agency. Moreover, Transparency goes beyond technical features. It is an important requirement also for the processes of policy and lawmaking. In this way, EAD1e's Pillars form the holistic ethical grounding upon which the Principles can build, and the latter may apply in various spheres of human activity.

EAD1e Pillars Mapped to General Principles

		EAD Pillars		
		Universal Human Values	Political Self-Determination Data Agency	Technical Dependability
EAD General Principles	Human Rights	■	■	
	Well-being	■	■	
	Data Agency	■	■	■
	Effectiveness			■
	Transparency	■	■	■
	Accountability	■	■	■
	Awareness of Misuse			■
	Competence			■

■ Indicates General Principle mapped to Pillar.

Mapping the Principles to the Content of the Chapters

The Chapters of *Ethically Aligned Design* provide in-depth subject matter expertise that allows readers to move from the General Principles to more deeply analyze ethical A/IS issues within the context of their specific work.

The mapping or indexing provided in the table below serve as directional starting points since elements of a Principle like "Competence" may resonate in several EAD1e Chapters. In addition, where core subjects are primarily covered by specific Chapters, we have done our best to indicate this via our mapping below.

EAD1e General Principles Mapped to Chapters

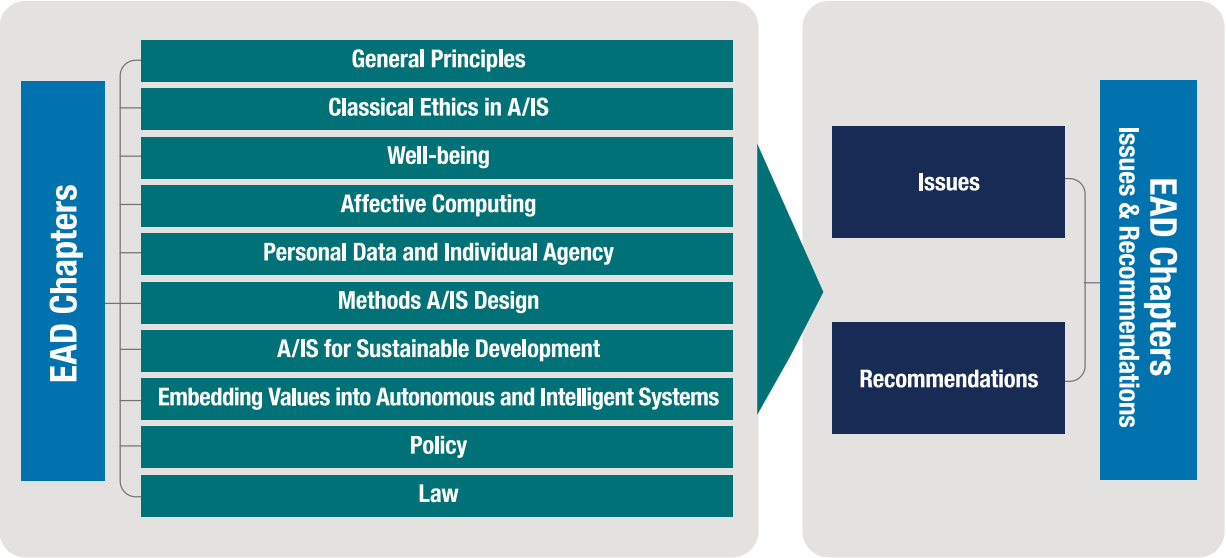
EAD General Principles	EAD Chapters: General Principles	Classical Ethics in A/IS	Well-being	Affective Computing	Data & Individual Agency	Methods A/IS Design	A/IS for Sustainable Dev.	Embedding Values into A/IS	Policy	Law
Human Rights	■	■	■	■	■	■	■	■	■	■
Well-being	■	■	■ ■	■	■		■	■	■	■
Data Agency	■		■	■	■ ■	■	■	■	■	
Effectiveness	■			■		■		■	■	■
Transparency	■			■		■		■	■	■
Accountability	■			■		■	■	■	■	■
Awareness of Misuse	■	■		■		■		■	■	■
Competence	■			■		■		■	■	■

■ Indicates General Principle mapped to Chapter.

■ Indicates primary EAD Chapter providing elaboration on a General Principle.

From Principles to Practice

It is at this step of the *Ethically Aligned Design* Conceptual Framework that readers will be able to identify the Principles and Chapters of key relevance to their work. Content provided in EAD1e Chapters is organized by "Issues" identified as the most pressing ethical matters surrounding A/IS design to address today and "Recommendations" on how it should be done. By reviewing these Issues and Recommendations in light of a specific A/IS product, service, or system being designed, readers are provided with a simple form of impact assessment and due diligence process to help put their "Principles into Practice" for themselves. Of course, more fine-tuned customization and adaptation of the content of EAD1e to fit specific sectors or applications are possible and will be pursued in the near future. See below for some implementation examples already happening.



Ethically Aligned Design in Implementation

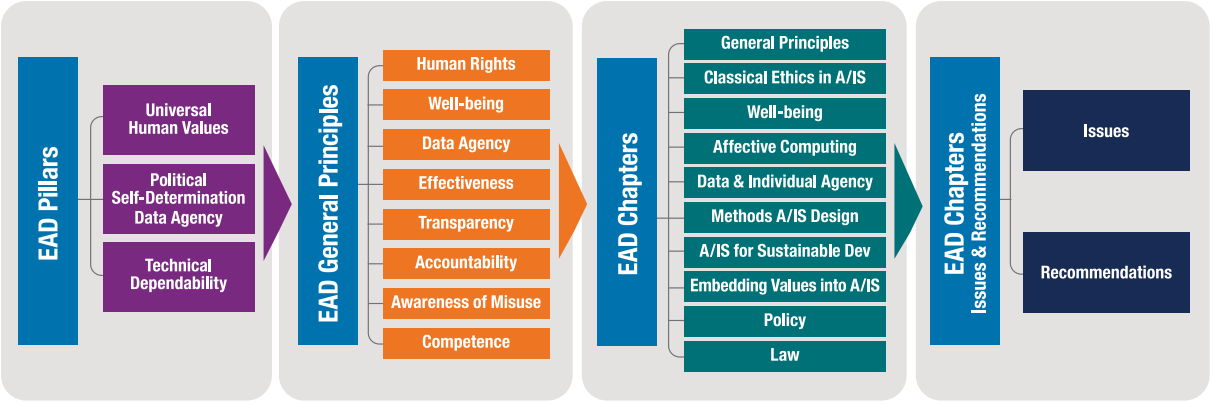
Ethically Aligned Design, *First Edition* represents the culmination of a three-year process guided bottom-up since 2015 by the rigor and standards of the engineering profession and by a globally open and iterative process involving hundreds of global experts. The analysis of the Principles, Issues, and Recommendations generated as part of an iterative process have already inspired the creation of fourteen IEEE Standardization Projects, a Certification Program, A/IS Ethics Courses, and multiple other action-oriented programs currently in development.

In its earlier manifestations, *Ethically Aligned Design* informed collaborations on A/IS governance with a broad range of governmental and civil society organizations, including the United Nations, the European Commission, the Organization for Economic Cooperation and Development and many national and municipal governments and institutions.[5] Moreover, the engagement in all of these arenas and with such partners has put the collective knowledge and creativity of The IEEE Global Initiative in the service of global policy-making with tangible and visible results. Beyond inspiring the policy arena, EAD1e and this growing body of work has also been influencing the development of industry-related resources.[6]

It is time to move "From Principles to Practice" in society regarding the governance of emerging autonomous and intelligent systems. The implementation of ethical principles must be validated by dependable applications of A/IS in practice while honoring our desire for political self-determination and data agency. To achieve societal progress, the autonomous and intelligent systems we create must be trustworthy, provable, and accountable and must align to our explicitly formulated human values.

It is our hope that *Ethically Aligned Design* and this conceptual framework will provide action-oriented inspiration for your work as well.

Ethically Aligned Design **Conceptual Framework—From Principles to Practice**



For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

Endnotes

[1] We prefer not to use—as far as possible—the vague term "AI" and use instead the term autonomous and intelligent systems (A/IS). This terminology is applied throughout *Ethically Aligned Design, First Edition* to ensure the broadest possible application of ethical considerations in the design of the addressed technologies and systems.

[2] See also *Draft Ethics Guidelines for Trustworthy AI* of The European Commission's High Level Expert Group on AI.

[3] A/IS should be subject to specific certification procedures by competent and qualified agencies with participation or control of public authorities in the same way other technical systems require certification before deployment. The IEEE has launched one of the world's first programs dedicated to creating A/IS certification processes. The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) offers processes by which organizations can seek certified A/IS products, systems, and services. It is being developed through an extensive and open public-private collaboration.

[4] For their overall framing, see the "General Principles" Chapter.

[5] As an example, the recently published report *Draft Ethics Guidelines for Trustworthy AI* of The European Commission's High Level Expert Group on AI explicitly mentions EAD as a major source of their inspiration. EAD has also been guiding policy creation for efforts of the United Nations and the Organization for Economic Cooperation and Development.

[6] Everyday Ethics for Artificial Intelligence: A Practical Guide for Designers and Developers

General Principles

The General Principles of *Ethically Aligned Design* articulate high-level ethical principles 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, in real, virtual, contextual, and mixed-reality environments.

The General Principles define imperatives for the design, development, deployment, adoption, and decommissioning of autonomous and intelligent systems. The Principles consider the role of A/IS creators, i.e., those who design and manufacture, of operators, i.e., those with expertise specific to use of A/IS, other users, and any other stakeholders or affected parties.

We have created these ethical General Principles for A/IS that:

- Embody the highest ideals of human beneficence within human rights.
- Prioritize benefits to humanity and the natural environment from the use of A/IS over commercial and other considerations. Benefits to humanity and the natural environment should not be at odds—the former depends on the latter. Prioritizing human well-being does not mean degrading the environment.
- Mitigate risks and negative impacts, including misuse, as A/IS evolve as socio-technical systems, in particular by ensuring actions of A/IS are accountable and transparent.

These General Principles are elaborated in subsequent sections of this chapter of *Ethically Aligned Design*, with specific contextual, cultural, and pragmatic explorations which impact their implementation.

General Principles as Imperatives

We offer high-level General Principles in *Ethically Aligned Design* that we consider to be imperatives for creating and operating A/IS that further human values and ensure trustworthiness. In summary, our General Principles are:

1. **Human Rights–**A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
2. **Well-being–**A/IS creators shall adopt increased human well-being as a primary success criterion for development.
3. **Data Agency–**A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
4. **Effectiveness–**A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
5. **Transparency–**The basis of a particular A/IS decision should always be discoverable.
6. **Accountability–**A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
7. **Awareness of Misuse–**A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
8. **Competence–**A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

Principle 1—Human Rights

A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.

Background

Human benefit is a crucial goal of A/IS, as is respect for human rights set out in works including, but not limited to: The Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights, the Convention on the Rights of the Child, the Convention on the Elimination of all forms of Discrimination against Women, the 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 principle that A/IS should be designed and operated in a way that both respects and fulfills human rights, freedoms, human dignity, and cultural diversity.

While their interpretation may change over time, "human rights", as defined by international law, provide a unilateral basis for creating any A/IS, as these systems 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 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 into account human rights in the way A/IS are developed, operated, tested, and validated. In short, human rights should be part of the ethical risk assessment of A/IS.

Recommendations

To best respect 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. Specifically:

- Governance frameworks, including standards and regulatory bodies, should be established to oversee processes which ensure that the use of A/IS does not infringe upon human rights, freedoms, dignity, and privacy, and which ensure traceability. This will contribute to building public trust in A/IS.
- A way to translate existing and forthcoming legal obligations into informed policy and technical considerations is needed. Such a method should allow for diverse cultural norms as well as differing legal and regulatory frameworks.

General Principles

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

Further Resources

The following documents and 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.
- N. Wiener, *The Human Use of Human Beings*, New York: Houghton Mifflin, 1954.
- 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, 1990.
- 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, 2018.
- The UN Guiding Principles on Business and Human Rights, 2011.
- British Standards Institute BS8611:2016, Robots and Robotic Devices. Guide to the Ethical Design and Application of Robots and Robotic Systems

Principle 2—Well-being

A/IS creators shall adopt increased human well-being as a primary success criterion for development.

Background

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

Well-being, for the purpose of *Ethically Aligned Design*, is based on the Organization for Economic Co-operation and Development's (OECD) "Guidelines on Measuring Subjective Well-being" perspective that, "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 as identified within Nassbaum-Sen's capability approach whereby well-being is objectively defined in terms of human capabilities necessary for functioning and flourishing.

Since modern societies will be largely constituted of A/IS users, we believe these considerations to be relevant for A/IS creators.

A/IS technologies can be narrowly conceived from an ethical standpoint. They can be legal, profitable, and safe in their usage, yet not positively contribute to human and environmental well-being. This means technologies created with the best intentions, but without considering well-being, 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.

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 Metric for Autonomous and Intelligent Systems.
- The Measurement of Economic Performance and Social Progress now commonly referred to as "The Stiglitz Report", commissioned by the then President of the French Republic, 2009. From the report: "…the time is ripe for our measurement system to shift emphasis from measuring economic production to measuring

General Principles

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."

- OECD Guidelines on Measuring Subjective Well-being, 2013.
- OECD Better Life Index, 2017.
- World Happiness Reports, 2012 – 2018.
- United Nations Sustainable Development Goal (SDG) Indicators, 2018.
- Beyond GDP, European Commission, 2018. From the site: "The Beyond GDP initiative is about developing indicators that are as clear and appealing as GDP, but more inclusive of environmental and social aspects of progress."
- Genuine Progress Indicator, State of Maryland (first developed by Redefining Progress), 2015.
- The International Panel on Social Progress, Social Justice, Well-Being and Economic Organization, 2018.
- R. Veenhoven, World Database of Happiness, Erasmus University Rotterdam, The Netherlands, Accessed 2018 at: http://worlddatabaseofhappiness.eur.nl.
- 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.

Principle 3—Data Agency

A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.

Background

Digital consent is a misnomer in its current manifestation. Terms and conditions or privacy policies are largely designed to provide legally accurate information regarding the usage of people's data to safeguard institutional and corporate interests, while often neglecting the needs of the people whose data they process. "Consent fatigue", the constant request for agreement to sets of long and unreadable data handling conditions, causes a majority of users to simply click and accept terms in order to access the services they wish to use. General obfuscation regarding privacy policies, and scenarios like the Cambridge Analytica scandal in 2018, demonstrate that even when individuals provide consent, the understanding of the value regarding their data and its safety is out of an individual's control.

This existing model of data exchange has eroded human agency in the algorithmic age. People don't know how their data is being used at all times or when predictive messaging is honoring their existing preferences or manipulating them to create new behaviors.

Regulations like the EU General Data Protection Regulation (GDPR) will help improve this lack of clarity regarding the exchange of personal data. But compliance with existing models of consent is not enough to safeguard people's agency regarding their personal information. In an era where A/IS are already pervasive in society, governments must recognize that limiting the misuse of personal data is not enough.

Society must also recognize that human rights in the digital sphere don't exist until individuals globally are empowered with means—including tools and policies—that ensure their dignity through some form of sovereignty, agency, symmetry, or control regarding their identity and personal data. These rights rely on individuals being able to make their choices, outside of the potential influence of biased algorithmic messaging or bad actors. Society also needs to be confident that those who are unable to provide legal informed consent, including minors and people with diminished capacity to make informed decisions, do not lose their dignity due to this.

Recommendation

Organizations, including governments, should immediately explore, test, and implement technologies and policies that let individuals specify their online agent for case-by-case authorization decisions as to who can process what personal data for what purpose. For minors and those with diminished capacity to make informed decisions, current guardianship approaches should be viewed to determine their suitability in this context.

General Principles

The general solution to give agency to the individual is meant to anticipate and enable individuals to own and fully control autonomous and intelligent (as in capable of learning) technology that can evaluate data use requests by external parties and service providers. This technology would then provide a form of "digital sovereignty" and could issue limited and specific authorizations for processing of the individual's personal data wherever it is held in a compatible system.

Further Resources

The following resources are designed to provide governments and other organizations—corporate, for-profit, not-for-profit, B Corp, or any form of public institution—basic information on services designed to provide user agency and/or sovereignty over their personal data.

- The European Data Protection Supervisor defines personal information management systems (PIMS) as:
- "...systems that help give individuals more control over their personal data...allowing individuals to manage their personal data in secure, local or online storage systems and share them when and with whom they choose. Providers of online services and advertisers will need to interact with the PIMS if they plan to process individuals' data. This can enable a human centric approach to personal information and new business models." For further information and ongoing research regarding PIMS, visit Crtl-Shift's PIMS monthly archive.

- IEEE P7006™, IEEE Standards Project for Personal Data Artificial Intelligence (AI) Agent describes the technical elements required to create and grant access to a personalized Artificial Intelligence that will comprise inputs, learning, ethics, rules, and values controlled by individuals.
- IEEE P7012™, IEEE Standards Project for Machine Readable Personal Privacy Terms is designed to provide individuals with a means to proffer their own terms respecting personal privacy in ways that can be read, acknowledged, and be agreed to by machines operated by others in the networked world.

Principle 4—Effectiveness

Creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.

Background

The responsible adoption and deployment of A/IS are essential if such systems are to realize their many potential benefits to the well-being of both individuals and societies. A/IS will not be trusted unless they can be shown to be effective in use. Harms caused by A/IS, from harm to an individual through to systemic damage, can undermine the perceived value of A/IS and delay or prevent its adoption.

Operators and other users will therefore benefit from measurement of the effectiveness of the A/IS in question. To be adequate, effective measurements need to be both valid and accurate, as well as meaningful and actionable. And such measurements must be accompanied by practical guidance on how to interpret and respond to them.

Recommendations

1. Creators engaged in the development of A/IS should seek to define metrics or benchmarks that will serve as valid and meaningful gauges of the effectiveness of the system in meeting its objectives, adhering to standards and remaining within risk tolerances. Creators building A/IS should ensure that the results when the defined metrics are applied are readily obtainable by all interested parties, e.g., users, safety certifiers, and regulators of the system.
2. Creators of A/IS should provide guidance on how to interpret and respond to the metrics generated by the systems.
3. To the extent warranted by specific circumstances, operators of A/IS should follow the guidance on measurement provided with the systems, i.e., which metrics to obtain, how and when to obtain them, how to respond to given results, and so on.
4. To the extent that measurements are sample-based, measurements should account for the scope of sampling error, e.g., the reporting of confidence intervals associated with the measurements. Operators should be advised how to interpret the results.
5. Creators of A/IS should design their systems such that metrics on specific deployments of the system can be aggregated to provide information on the effectiveness of the system across multiple deployments. For example, in the case of autonomous vehicles, metrics should be generated both for a specific instance of a vehicle and for a fleet of many instances of the same kind of vehicle.
6. In interpreting and responding to measurements, allowance should be made for variation in the specific objectives and circumstances of a given deployment of A/IS.

General Principles

7. To the extent possible, industry associations or other organizations, e.g., IEEE and ISO, should work toward developing standards for the measurement and reporting on the effectiveness of A/IS.

Further Resources

- R. Dillmann, KA 1.10 Benchmarks for Robotics Research, 2010.
- A. Steinfeld, T.W. Fong, D. Kaber, J. Scholtz, A. Schultz, and M. Goodrich, "Common Metrics for Human-Robot Interaction", 2006 Human-Robot Interaction Conference, March, 2006.
- R. Madhavan, E. Messina, and E. Tunstel, Eds., Performance Evaluation and Benchmarking of Intelligent Systems, Boston, MA: Springer, 2009.
- *IEEE Robotics & Automation Magazine*, Special Issue on Replicable and Measurable Robotics Research, Volume 22, No. 3, September 2015.
- C. Flanagin, A Survey on Robotics Systems and Performance Analysis, 2011.
- Transaction Processing Performance Council (TPC) Establishes Artificial Intelligence Working Group (TPC-AI) tasked with developing industry standard benchmarks for both hardware and software platforms associated with running Artificial Intelligence (AI) based workloads, 2017.

Principle 5—Transparency

The basis of a particular A/IS decision should always be discoverable.

Background

A key concern over autonomous and intelligent 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. 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. The term "transparency" in the context of A/IS also addresses the concepts of traceability, explainability, and interpretability.

A/IS will perform tasks that are far more complex and have more effect on our world than prior generations of technology. Where the task is undertaken in a non-deterministic manner, it may defy simple explanation. 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 or driverless car autopilots. Systems such as these are safety-critical systems.

At the same time, the complexity of A/IS technology and the non-intuitive way in which it may operate will make it difficult for users of those systems to understand the actions of the A/IS that they use, or with which they interact. This opacity, combined with the often distributed manner in which the A/IS are developed, will complicate efforts to determine and allocate responsibility when something goes wrong. Thus, lack of transparency increases the risk and magnitude of harm when users do not understand the systems they are using, or there is a failure to fix faults and improve systems following accidents. Lack of transparency also increases the difficulty of ensuring accountability (see Principle 6— Accountability).

Achieving transparency, which may involve a significant portion of the resources required to develop the A/IS, is important to each stakeholder group for the following reasons:

1. For users, what the system is doing and why.
2. For creators, including those undertaking the validation and certification of A/IS, the systems' processes and input data.
3. For an accident investigator, if accidents occur.
4. For those in the legal process, to inform evidence and decision-making.
5. For the public, to build confidence in the technology.

General Principles

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, including but not limited to, the following use cases:

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.
3. For accident investigators, secure storage of sensor and internal state data comparable to a flight data recorder or black box.

IEEE P7001™, IEEE Standard for Transparency of Autonomous Systems is one such standard, developed in response to this recommendation.

Further Resources

- C. Cappelli, 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, pp. 1–17, Jouy-en-Josas, France: HEC Paris, 2013.
- J.C. Sampaio do Prado Leite and C. Cappelli, "Software Transparency." *Business & Information Systems Engineering* 2, no. 3, pp. 127–139, 2010.
- A, Winfield, and M. Jirotka, "The Case for an Ethical Black Box," *Lecture Notes in Artificial Intelligence* 10454, pp. 262–273, 2017.
- R. R. Wortham, 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.
- M. Scherer, "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," Report of the U.K. House of Commons Science and Technology Committee on Robotics and Artificial Intelligence, pp. 17-18, September 13, 2016.

Principle 6—Accountability

A/IS shall be created and operated to provide an unambiguous rationale for decisions made.

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 to avoid potential harm. Additionally, manufacturers of these systems must be accountable in order to address legal issues of culpability. It should, if necessary, be possible to apportion culpability among responsible creators (designers and manufacturers) and operators to avoid confusion or fear within the general public.

Accountability and partial accountability are not possible without transparency, thus this principle is closely linked with Principle 5–Transparency.

Recommendations

To best address issues of responsibility and accountability:

1. Legislatures/courts should clarify responsibility, culpability, liability, and accountability for A/IS, where possible, prior to 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, the diversity of existing cultural norms among the groups of users of these A/IS.

3. Multi-stakeholder ecosystems including creators, and government, civil, and commercial stakeholders, should be developed to help establish norms where they do not exist because A/IS-oriented technology and their impacts are too new. These ecosystems would include, but not be limited to, representatives of civil society, law enforcement, insurers, investors, manufacturers, engineers, lawyers, and users. The norms can mature into best practices and laws.

General Principles

4. Systems for registration and record-keeping should be established so that it is always possible to find out who is legally responsible for a particular A/IS. Creators, including manufacturers, along with operators, of A/IS should register key, high-level parameters, including:
    - Intended use,
    - Training data and training environment, if applicable,
    - Sensors and real world data sources,
    - Algorithms,
    - Process graphs,
    - Model features, at various levels,
    - User interfaces,
    - Actuators and outputs, and
    - Optimization goals, loss functions, and reward functions.

Further Resources

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

Principle 7—Awareness of Misuse

Creators shall guard against all potential misuses and risks of A/IS in operation.

Background

New technologies give rise to greater risk of deliberate or accidental misuse, and this is especially true for A/IS. A/IS increases the impact of risks such as hacking, misuse of personal data, system manipulation, or exploitation of vulnerable users by unscrupulous parties. Cases of A/IS hacking have already been widely reported, with driverless cars, for example. The Microsoft Tay AI chatbot was famously manipulated when it mimicked deliberately offensive users. In an age where these powerful tools are easily available, there is a need for a new kind of education for citizens to be sensitized to risks associated with the misuse of A/IS. The EU's General Data Protection Regulation (GDPR) provides measures to remedy the misuse of personal data.

Responsible innovation requires A/IS creators to anticipate, reflect, and engage with users of A/IS. Thus, citizens, lawyers, governments, etc., all have a role to play through education and awareness in developing accountability structures (see Principle 6), in addition to guiding new technology proactively toward beneficial ends.

Recommendations

1. Creators should be aware of methods of misuse, and they should design A/IS in ways to minimize the opportunity for these.
2. Raise public awareness around the issues of potential A/IS technology misuse in an informed and measured way by:
    - Providing ethics education and security awareness that sensitizes society to the potential risks of misuse of A/IS. For example, provide "data privacy warnings" that some smart devices will collect their users' personal data.
    - Delivering this education in scalable and effective ways, including having experts with the greatest credibility and impact who can minimize unwarranted fear about A/IS.
    - Educating government, lawmakers, and enforcement agencies about these issues of A/IS so citizens can work collaboratively with these agencies to understand safe use of A/IS. For example, the same way police officers give public safety lectures in schools, they could provide workshops on safe use and interaction with A/IS.

Further Resources

- A. Greenberg, "Hackers Fool Tesla S's Autopilot to Hide and Spoof Obstacles," Wired, August 2016.
- C. Wilkinson and E. Weitkamp, *Creative Research and Communication: Theory and Practice*, Manchester, UK: Manchester University Press, 2016 (in relation to Recommendation #2).
- Engineering and Physical Sciences Research Council, "Anticipate, Reflect, Engage and Act (AREA)," Framework for Responsible Research and Innovation, Accessed 2018.

Principle 8—Competence

Creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

Background

A/IS can and often do make decisions that previously required human knowledge, expertise, and reason. Algorithms potentially can make even better decisions, by accessing more information, more quickly, and without the error, inconsistency, and bias that can plague human decision-making. As the use of algorithms becomes common and the decisions they make become more complex, however, the more normal and natural such decisions appear.

Operators of A/IS can become less likely to question and potentially less able to question the decisions that algorithms make. Operators will not necessarily know the sources, scale, accuracy, and uncertainty that are implicit in applications of A/IS. As the use of A/IS expands, more systems will rely on machine learning where actions are not preprogrammed and that might not leave a clear record of the steps that led the system to its current state. Even if those records do exist, operators might not have access to them or the expertise necessary to decipher those records.

Standards for the operators are essential. Operators should be able to understand how A/IS reach their decisions, the information and logic on which the A/IS rely, and the effects of those decisions. Even more crucially, operators should know when they need to question A/IS and when they need to overrule them.

Creators of A/IS should take an active role in ensuring that operators of their technologies have the knowledge, experience, and skill necessary not only to use A/IS, but also to use it safely and appropriately, towards their intended ends. Creators should make provisions for the operators to override A/IS in appropriate circumstances.

While standards for operator competence are necessary to ensure the effective, safe, and ethical application of A/IS, these standards are not the same for all forms of A/IS. The level of competence required for the safe and effective operation of A/IS will range from elementary, such as "intuitive" use guided by design, to advanced, such as fluency in statistics.

Recommendations

1. Creators of A/IS should specify the types and levels of knowledge necessary to understand and operate any given application of A/IS. In specifying the requisite types and levels of expertise, creators should do so for the individual components of A/IS and for the entire systems.

2. Creators of A/IS should integrate safeguards against the incompetent operation of their systems. Safeguards could include issuing

General Principles

notifications/warnings to operators in certain conditions, limiting functionalities for different levels of operators (e.g., novice vs. advanced), system shut-down in potentially risky conditions, etc.

3. Creators of A/IS should provide the parties affected by the output of A/IS with information on the role of the operator, the competencies required, and the implications of operator error. Such documentation should be accessible and understandable to both experts and the general public.

4. Entities that operate A/IS should create documented policies to govern how A/IS should be operated. These policies should include the real-world applications for such A/IS, any preconditions for their effective use, who is qualified to operate them, what training is required for operators, how to measure the performance of the A/IS, and what should be expected from the A/IS. The policies should also include specification of circumstances in which it might be necessary for the operator to override the A/IS.

5. Operators of A/IS should, before operating a system, make sure that they have access to the requisite competencies. The operator need not be an expert in all the pertinent domains but should have access to individuals with the requisite kinds of expertise.

Further Resources

- S. Barocas and A.D. Selbst, The Intuitive Appeal of Explainable Machines, Fordham Law Review, 2018.
- W. Smart, C. Grimm, and W. Hartzog, "An Education Theory of Fault for Autonomous Systems", 2017.

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The General Principles Committee

- **Alan Winfield** (Founding Chair) – Professor, Bristol Robotics Laboratory, University of the West of England; Visiting Professor, University of York
- **Mark Halverson** (Co-Chair) – Founder and CEO at Precision Autonomy
- **Peet van Biljon** (Co-Chair) – Founder and CEO at BMNP Strategies LLC, advisor on strategy, innovation, and business transformation; Adjunct professor at Georgetown University; Business ethics author
- **Shahar Avin** – Research Associate, Centre for the Study of Existential Risk, University of Cambridge
- **Bijilash Babu** – Senior Manager, Ernst and Young, EY Global Delivery Services India LLP
- **Richard Bartley** – Senior Director - Analyst, Security & Risk Management, Gartner, Toronto, Canada Security Principal Director, Accenture, Toronto, Canada.
- **R. R. Brooks** – Professor, Holcombe Department of Electrical and Computer Engineering, Clemson University
- **Nicolas Economou** – Chief Executive Officer, H5; Chair, Science, Law and Society Initiative at The Future Society Chair, Law Committee, Global Governance of AI Roundtable; Member, Council on Extended Intelligence (CXI)
- **Hugo Giordano** – Engineering Student at Texas A&M University
- **Alexei Grinbaum** – Researcher at CEA (French Alternative Energies and Atomic Energy Commission) and Member of the French Commission on the Ethics of Digital Sciences and Technologies CERNA
- **Jia He** – Independent Researcher, Graduate Delft University of Technology in Engineering and Public Policy, project member within United Nations, ICANN, and ITU Executive Director of Toutiao Research (Think Tank), Bytedance Inc.
- **Bruce Hedin** – Principal Scientist, H5
- **Cyrus Hodes** – Advisor AI Office, UAE Prime Minister's Office, Co-founder and Senior Advisor, AI Initiatives@The Future Society; Member, AI Expert Group at the OECD, Member, Global Council on Extended Intelligence; Co-founder and Senior Advisor, The AI Initiative @ The Future Society
- **Nathan F. Hutchins** – Applied Assistant Professor, Department of Electrical and Computer Engineering, The University of Tulsa
- **Narayana GPL. Mandaleeka ("MGPL")** – Vice President & Chief Scientist, Head, Business Systems & Cybernetics Centre, Tata Consultancy Services Ltd.
- **Vidushi Marda** – Programme Officer, ARTICLE 19
- **George T. Matthew** – Chief Medical Officer, North America, DXC Technology

General Principles

- **Nicolas Miailhe** – Co-Founder & President, The Future Society; Member, AI Expert Group at the OECD; Member, Global Council on Extended Intelligence; Senior Visiting Research Fellow, Program on Science Technology and Society at Harvard Kennedy School. Lecturer, Paris School of International Affairs (Sciences Po); Visiting Professor, IE School of Global and Public Affairs
- **Rupak Rathore** – Principal Consultant at ATCS for Telematics, Connected Car and Internet of Things; Advisor on strategy, innovation and transformation journey management; Senior Member, IEEE
- **Peter Teneriello** – Investment Analyst, Private Equity and Venture Capital, TMRS
- **Niels ten Oever** – Head of Digital, Article 19, Co-chair Research Group on Human Rights Protocol Considerations in the Internet Research Taskforce (IRTF)
- **Alan R. Wagner** – Assistant Professor, Department of Aerospace Engineering, Research Associate, The Rock Ethics Institute, The Pennsylvania State University.

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

Classical Ethics in A/IS

We applied 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, we drew from classical ethics theories and the disciplines of machine ethics, information ethics, and technology ethics.

As direct control over tools becomes further removed, creators of autonomous systems must ask themselves how cultural and ethical presumptions bias artificially intelligent creations. Such introspection is more necessary than ever because the precise and deliberate design of algorithms in self-sustained digital systems will result in responses based on such design.

By drawing from over two thousand years' worth of classical ethics traditions, we explore established ethics systems, including both philosophical traditions (utilitarianism, virtue ethics, and deontological ethics) and religious and culture-based ethical systems (Buddhism, Confucianism, African Ubuntu traditions, and Japanese Shinto) and their stance on human morality in the digital age.[1] In doing so, we critique assumptions around concepts such as good and evil, right and wrong, virtue and vice, and we attempt to carry these inquiries into artificial systems' decision-making processes.

Through reviewing the philosophical foundations that define autonomy and ontology, we 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, we address notions of responsibility and accountability for the decisions made by autonomous systems and other artificially intelligent technologies.

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 formation of the individual character (*ethos*) is intrinsically related to the others, as well as to the tasks of administration of work within the family (*oikos*). Eventually, this all expands into the framework of the *polis*, or public space (*poleis*). When we discuss ethical issues of A/IS, it becomes crucial to consider these three traditional economic dimensions, since western classical ethics was developed from this foundation and has evolved in modernity into an individual morality disconnected from economics and politics. This disconnection has been questioned and explored by thinkers such as Adam Smith, Georg W. F. Hegel, Karl 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 it. The moral autonomous subject of modernity became thus a worldless isolated subject. This process is important to understand in terms of ethics for A/IS since it is, paradoxically, the kind of autonomy that is supposed to be achieved by intelligent machines as humans evolve 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. While the concepts of artificial intelligence and autonomy are mainly used metaphorically as technical terms in computer science and technology, general and popular discourse may not share in the same nuanced understanding, and political and societal discourse may become distorted or

Classical Ethics in A/IS

misleading. The question of whether A/IS and the terminology used to describe them will have any kind of impact on our conception of autonomy depends on our policy toward it. For example, the commonly held fear that A/IS will relegate humanity to mere spectators or slaves, whether realistic or not, is informed by our view of, and terminology around, A/IS. Such attitudes are flexible and can be negotiated. As noted above, present human societies are being redefined in terms of digital citizenship via online 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.

There is still value that can be gained by considering how Western ethical traditions can be integrated into either A/IS public awareness campaigns or supplemented in engineering and science education programs, as noted under the issue "Presenting ethics to the creators of A/IS". Below is a short overview of how four different traditions can add value.

- **Virtue ethics:** Aristotle argues, using the concept of *telos*, or goal, that the ultimate goal of humans is "eudaimonia", roughly translated as "flourishing". A moral agent achieves "flourishing"—since it is an action, not a state—by constantly balancing factors including social environment, material provisions, friends, family, and one's own self. One cultivates the self through habituation, practicing and strengthening virtuous action as the "golden mean" (a principle of rationality). Such cultivation requires an appropriate balance between extremes of excess and deficiency, which Aristotle identifies as vices. In the context of A/IS, virtue ethics has two immediate values. First, it provides a model for iterative learning and growth, and moral value informed by context and practice, not just as compliance with a given, static ruleset. Second, it provides to those who develop and implement A/IS a framework to counterbalance tendencies toward excess, which are common in economically-driven environments.

- **Deontological ethics:** As developed by 18th century German philosopher, Immanuel Kant, the basic premise of deontological ethics addresses the concept of duty. Humans have a rational capacity to create and abide by rules that allow for duty-based ethics to emerge. Rules that produce duties are said to have value in themselves without requiring a greater-good justification. Such rules are fundamental to our existence, self-worth, and to creating conditions that allow for peaceful coexistence and interaction, e.g., the duty not to harm others; the duty not to steal. To identify rules that can be universalized and made duties, Kant uses the categorical imperative: "Act only on that maxim through which you can at the same time will that it should become a universal law." This means the rule must be inherently desirable, doable, valuable, and others must be able to understand and follow it. Rules based merely on personal choice without wider appeal are not capable of universalization. There is mutual reciprocity in rule-making and rule adherence; if you "will" that a rule should become universal law, you not only contribute

Classical Ethics in A/IS

to rule creation but also agree to be bound by the same rule. The rule should be action-guiding, i.e., recommending, prescribing, limiting, or proscribing action. Kant also uses the humanity formulation of the categorical imperative: "Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end." This produces duties to respect humanity and human dignity, and not to treat either as a means to an end.

- In the context of A/IS, one consideration is to wonder if developers are acting with the best interests of humanity and human dignity in mind. This could possibly be extended to A/IS whereby they are assisting humanity as an instrument of action that has an impact on decision-making capabilities, despite being based on neural machine learning or set protocols. The humanity formulation of "the categorical imperative" has implications for various scenarios. The duty to respect human dignity may require some limitations on the functions and capability of A/IS so that they do not completely replace humans, human functions, and/or "human central thinking activities" such as judgment, discretion, and reasoning. Privacy and safeguarding issues around A/IS assisting humans, e.g., healthcare robots, may require programming certain values so that A/IS do not divulge personal information to third parties, or compromise a human's physical or mental well-being. It may also involve preventing A/IS from deceiving or manipulating humans.
- Potential benefits and financial incentives from exploiting A/IS may provide ends-means justifications for their use, while disregarding the treatment of humanity as an end in itself, e.g., cutting back on funding rigorous testing of A/IS before they reach the market and society. Maintaining human agency in human-machine interaction is a manifestation of the duty to respect human dignity. For example, a human has the right to know when they are interacting with A/IS, and may require consent for any A/IS interaction.
- **Utilitarian ethics:** Also called consequentialist ethics, this code of ethics refers to the consequences of one's decisions and actions. According to the utility principle, the right course of action is the one that maximizes the utility (utilitarianism) or pleasure (hedonism) for the greatest number of people. This ethics theory does, however, warn against superficial and short-term evaluations of utility or pleasure. Therefore, it is the responsibility of the A/IS developers to consider long-term effects. Social justice is paramount in this instance, thus it must be ascertained if the implementation of A/IS will contribute to humanity, or negatively impact employment or other capabilities. Indeed, where it is deemed A/IS can supplement humanity, it should be designed in such a way that the benefits are obvious to all the stakeholders.
- **Ethics of care:** Generally viewed as an instance of feminist ethics, this approach emphasizes the importance of relationships which is context-bound. Relationships are ontologically basic to humanity, according to Nel Noddings, feminist and philosopher of education; to care for other human beings is one of our basic human attributes. For such

Classical Ethics in A/IS

a theory to have relevance in this context, one needs to consider two criteria: 1) the relationship with the other person, or entity, must already exist or must have the potential to exist, and 2) the relationship should have the potential to grow into a caring relationship. Applied to A/IS, an interesting question comes to the foreground: Can one care for humans and their interests in tandem with non-human entities? If one expects A/IS to be beneficial to humanity, as in the instance of robots assisting with care of the elderly, then can one deduce the possibility of humans caring for A/IS? If that possibility exists, do principles of social justice become applicable to A/IS?

Recommendations

By returning to classical ethics foundations, expand the discussion on ethics in 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. They move according to their programming, following rules that are designed by humans to be moral.

Expand 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

- J. Bielby, Ed., "Digital Global Citizenship," International Review of Information Ethics, vol. 23, pp. 2-3, Nov. 2015.
- O. Bendel, "Towards Machine Ethics," in Technology Assessment and Policy Areas of Great Transitions: Proceedings from the PACITA 2013 Conference in Prague, PACITA 2013, Prague, March 13-15, 2013, T. Michalek, L. Hebáková, L. Hennen, C. Scherz, L. Nierling, J. Hahn, Eds. Prague: Technology Centre ASCR, 2014. pp. 321-326.
- O. Bendel, "Considerations about the Relationship between Animal and Machine Ethics," AI & Society, vol. 31, no. 1, pp. 103-108, Feb. 2016.
- N. Berberich and K. Diepold, "The Virtuous Machine - Old Ethics for New Technology?" arXiv:1806.10322 [cs.AI], June 2018.
- R. Capurro, M. Eldred, and D. Nagel, Digital Whoness: Identity, Privacy and Freedom in the Cyberworld. Berlin: Walter de Gruyter, 2013.
- D. Chalmers, "The Singularity: A Philosophical Analysis," Journal of Consciousness Studies, vol. 17, pp. 7-65, 2010.
- D. Davidson, "Representation and Interpretation," in Modelling the Mind, K. A. M. Said, W. H. Newton-Smith, R. Viale, and K. V. Wilkes, Eds. New York: Oxford University Press, 1990, pp. 13-26.
- N. Noddings, Caring: A Relational Approach to Ethics and Moral Education. Oakland, CA: University of California Press, 2013.
- O. Ulgen, "Kantian Ethics in the Age of Artificial Intelligence and Robotics," QIL, vol. 43, pp. 59-83, Oct. 2017.

Classical Ethics in A/IS

- O. Ulgen, "The Ethical Implications of Developing and Using Artificial Intelligence and Robotics in the Civilian and Military Spheres," House of Lords Select Committee, Sept. 6, 2017, UK.
- O. Ulgen, "Human Dignity in an Age of Autonomous Weapons: Are We in Danger of Losing an 'Elementary Consideration of Humanity'?" in How International Law Works in Times of Crisis, I. Ziemele and G. Ulrich, Eds. Oxford: Oxford University Press, 2018.

Issue: The Distinction between Agents and Patients

Background

Of particular concern when understanding the relationship between human beings and A/IS is the uncritically applied anthropomorphic approach toward A/IS that many industry and policymakers 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 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 to implant true morality and emotions, and thus accountability, i.e., autonomy, into A/IS blurs the distinction between agents and patients and may encourage anthropomorphic 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 analyzing 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, the ethical critique of those customs, and the law are important distinctions.

The above misunderstanding in definitions of autonomy arises in part because of the tendency for humans to shape artificial creations in their own image, and our desire to lend our human

Classical Ethics in A/IS

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 begin to resemble human beings more closely. It is possible for terms like "artificial intelligence" or "morality of machines" to be used as metaphors without resulting in misunderstanding. 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; they 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 overcome general assumptions of anthropomorphism in A/IS and thereby address the uncertainty therein (Mill, 1999).

Recommendations

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 address general assumptions of anthropomorphism in A/IS.

Further Resources

- R. Capurro, "Toward a Comparative Theory of Agents." AI & Society, vol. 27, no. 4, pp. 479-488, Nov. 2012.
- W. J. King and J. Ohya, "The Representation of Agents: Anthropomorphism, Agency, and Intelligence," in Conference Companion on Human Factors in Computing Systems. Vancouver: ACM, 1996, pp. 289-290.
- W. Hofkirchner, "Does Computing Embrace Self-Organisation?" in Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation, G. Dodig-Crnkovic and M. Burgin, Eds. London: World Scientific, 2011, pp. 185-202.
- International Center for Information Ethics, 2018.
- J. S. Mill, On Liberty. London: Longman, Roberts & Green, 1869.
- P. P. Verbeek, What Things Do: Philosophical Reflections on Technology, Agency, and Design. University Park, PA: Pennsylvania State University Press, 2005.

Classical Ethics in A/IS

Issue: The 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 A/IS, 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-disciplinary instances, i.e., a conversation between 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 will go a long way in supporting a constructive dialogue on ethics and A/IS. With access and accessibility concerns intricately linked with education in communities, as well as secondary and tertiary institutions, society needs to take a vested interest in creating awareness 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.

Identifying terms that will be intelligible to all relevant audiences is pragmatic, but care should be taken not to dilute or misrepresent concepts that are familiar to moral philosophy and ethics. One way around this is to engage in applied ethics; illustrate how a particular concept would work in the A/IS context or scenario. Another way is to understand whether terminology used across different disciplines actually has the same or similar meaning and effect which can be expressed accordingly.

Recommendations

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-disciplinary exchanges whereby philosophy scholars and ethicists attend and present in non-philosophical courses. This will both raise awareness and sensitize non-philosophical scholars and practitioners to the vocabulary.

Further Resources

- R. T. Ames, Confucian Role Ethics: A Vocabulary. Hong Kong: Chinese University Press, 2011.
- R. Capurro, "Towards an Ontological Foundation of Information Ethics," Ethics and Information Technology, vol. 8, no. 4, pp. 175-186, 2006.
- S. Mattingly-Jordan, R. Day, B. Donaldson, P. Gray, and L. M. Ingram, "Ethically Aligned Design, First Edition Glossary," Prepared for The IEEE Global Initiative for Ethically Aligned Design, Feb. 2019.

Classical Ethics in A/IS

- B. M. Lowe, Emerging Moral Vocabularies: The Creation and Establishment of New Forms of Moral and Ethical Meanings. Lanham, MD: Lexington Books, 2006.
- D. J. Flinders, "In Search of Ethical Guidance: Constructing a Basis for Dialogue," International Journal of Qualitative Studies in Education, vol. 5, no. 2, pp. 101-115, 1992.
- G. S. Saldanha, "The Demon in the Gap of Language: Capurro, Ethics and Language in Divided Germany," in Information Cultures in the Digital Age. Wiesbaden, Germany: Springer Fachmedien, 2016, pp. 253-268.
- J. Van Den Hoven and G. J. Lokhorst, "Deontic Logic and Computer-Supported Computer Ethics," Metaphilosophy, vol. 33, no. 3, pp. 376-386, April 2002.

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. Keeping in mind that the task of philosophical ethics should be to examine good and evil, ethics should examine values, not prescribe them. Laws, which arise from ethics, are entrenched mores that have been critically assessed to prescribe.

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 inventions further entrenches ethical design practices. In the abstract, this is not so difficult to describe, but is very difficult to encode into systems. This problem can be addressed by providing students with job aids such as checklists, flowcharts, and matrices that will 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 didactic strategy which proves effective in artificial intelligence ethics education.

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 are 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.

Classical Ethics in A/IS

- 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 is referred to in the computer programming lexicon 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.

This further underscores the importance of education and training on ethical considerations relating to A/IS. Such courses should be developed and presented to students of engineering, A/IS, computer science, and other relevant fields. These courses do not add value *a posteriori,* but should be embedded from the beginning to allow for absorption of the underlying ethical considerations as well as allowing for critical thinking to come to fruition once the students graduate. There are various approaches that can be considered on a tertiary level:

- Parallel (information) ethics program that is presented together with the science program during the course of undergraduate and postgraduate study;
- Embedded (information) ethics modules within the science program, i.e., one module per semester;
- Short (information) ethics courses specifically designed for the science program that can be attended by the current students, alumni, or professionals. These will function as either introductory, refresher, or specialized courses.

Courses can also be blended to include students and/or practitioners from diverse backgrounds rather than the more traditional practice of homogenous groups, such as engineering students, continuing education programs directed at a specific specialization, and the like.

Recommendations

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. Curricula development is crucial in each approach. In addition to research articles and best practices, it is recommended that engineers and practitioners come together with social scientists and philosophers to develop case studies, interactive virtual reality gaming, and additional course interventions that are relevant to students.

Further Resources

- T. W. Bynum and S. Rogerson, Computer Ethics and Professional Responsibility. Malden, MA: Wiley-Blackwell, 2003.
- E. G. Seebauer and R. L. Barry, Fundamentals of Ethics for Scientists and Engineers. New York: Oxford University Press, 2001.

Classical Ethics in A/IS

- C. Whitbeck, "Teaching Ethics to Scientists and Engineers: Moral Agents and Moral Problems," Science and Engineering Ethics, vol. 1, no. 3, pp. 299-308, Sept. 1995.
- B. Zevenbergen, et al. "Philosophy Meets Internet Engineering: Ethics in Networked Systems Research," GTC Workshop Outcomes Paper. Oxford: Oxford Internet Institute, University of Oxford, 2015.
- M. Alvarez, "Teaching Information Ethics," International Review of Information Ethics, vol. 14, pp. 23-28, Dec. 2010.
- P. P. Verbeek, Moralizing Technology: Understanding and Designing the Morality of Things. Chicago, IL: University of Chicago Press, 2011.
- K. A. Joyce, K. Darfler, D. George, J. Ludwig, and K. Unsworth, "Engaging STEM Ethics Education," Engaging Science, Technology, and Society, vol. 4, no. 1-7, 2018.

Issue: Accessing Classical Ethics by Corporations and Companies

Background

Many companies, from startups to tech giants, understand that ethical considerations in tech design are increasingly important, but are not 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? Corporate workshops and exercises will need to go beyond opinion-gathering exercises to embed ethical considerations into structures, environments, training, and development.

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; there is also 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. Such an understanding includes that of the philosophical language used in ethics and the translation of that language across disciplines.

If we take, for instance, the terminology and usage of the concept of "trust" in reference to technology, the term "trust" has specific philosophical, legal, and engineering connotations. It is not an abstract concept. It is attributable to humans, and relates to claims and actions people make. Machines, robots, and algorithms lack the ability to make claims and so cannot be attributed with trust. They cannot determine whether something is trustworthy or not. Software engineers may refer to "trusting" the data, but this relates to the data's authenticity and veracity to ensure software performance. In the context of A/IS, "trust" means "functional reliability"; it means there is confidence in the technology's predictability, reliability, and security against hackers or impersonators of authentic users.

Classical Ethics in A/IS

Recommendations

In order to achieve multicultural, multidisciplinary, and multi-sectoral dialogues between technologists, philosophers, and policymakers, a nuanced understanding in philosophical and technical language, which is critical to digital society from Internet of Things (IoT), privacy, and cybersecurity to issues of Internet governance, must be translated into norms and made available to technicians and policymakers who may not understand the nuances of the terminology in philosophical, legal, and engineering contexts. It is therefore recommended that the translation of the critical-thinking terminology of philosophers, policymakers, and other stakeholders on A/IS be translated into norms accessible to technicians.

Further Resources

- A. Bhimani, "Making Corporate Governance Count: The Fusion of Ethics and Economic Rationality," Journal of Management & Governance, vol. 12, no. 2, pp. 135-147, June 2008.
- A. B. Carroll, "A History of Corporate Social Responsibility," in The Oxford Handbook of Corporate Social Responsibility, A. Chrisanthi, R. Mansell, D. Quah, and R. Silverstone, Eds. Oxford, U.K.: Oxford University Press, 2008.
- W. Lazonick, "Globalization of the ICT Labor Force," in The Oxford Handbook of Information and Communication Technologies, A. Chrisanthi, R. Mansell, D. Quah, and R. Silverstone, Eds. Oxford, U.K.: Oxford University Press, 2006.
- IEEE P7000™, IEEE Standards Project for Model Process for Addressing Ethical Concerns During System Design will provide engineers and technologists with an implementable process aligning innovation management processes, IT system design approaches, and software engineering methods to minimize ethical risk for their organizations, stakeholders and end users.

Issue: The 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 are increasing, where decisions made via aggregate algorithms directly impact employment prospects. The uncritical use of A/IS in the workplace, and its impact on employee-employer relations, is of utmost concern due to the high chance of error and biased outcome.

The concept of responsible research and innovation (RRI)is a growing area, particularly within the EU. It offers potential solutions to workplace bias and is being adopted by several research funders, such as the Engineering and Physical Sciences Research Council (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, from the design stage onwards.

Classical Ethics in A/IS

Quoting Rene Von Schomberg, science and technologies studies specialist and philosopher, "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)."**[2]

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.

This emerging and exciting new concept aims to also push the boundaries to incorporate relevant stakeholders whose influence in responsible research is on a global stage. While this concept initially focuses on the workplace setting, success will only be achieved through the active involvement from private companies of industry, AI Institutes, and those who are at the forefront in A/IS design. Responsible research and innovation will be achieved through careful research and innovation governance that will ensure research purpose, process, and outcomes that are acceptable, sustainable, and even desirable. It will be incumbent on RRI experts to engage at a level where private companies will feel empowered and embrace this concept as both practical to implement and enact.

Recommendations

It is recommended, through the application of RRI as founded in classical ethics theory, that 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

- M. Burget, E. Bardone, and M. Pedaste, "Definitions and Conceptual Dimensions of Responsible Research and Innovation: A Literature Review," Science and Engineering Ethics, vol. 23, no. 1, pp. 1-9, 2016.
- European Commission Communication, "Artificial Intelligence for Europe," COM 237, April, 2018.
- R. Von Schomberg, "Prospects for Technology Assessment in a Framework of Responsible Research and Innovation," in Technikfolgen Abschätzen Lehren: Bildungspotenziale Transdisziplinärer Methode. Wiesbaden, Germany: Springer VS, 2011, pp. 39-61.
- B. C. Stahl, G. Eden, M. Jirotka, M. Coeckelbergh, "From Computer Ethics to Responsible Research and Innovation in ICT: The Transition of Reference Discourses Informing Ethics-Related Research in Information Systems," Information & Management, vol. 51, no. 6, pp. 810-818, September 2014.

- B. C. Stahl, M. Obach, E. Yaghmaei, V. Ikonen, K. Chatfield, and A. Brem, "The Responsible Research and Innovation (RRI) Maturity Model: Linking Theory and Practice," Sustainability, vol. 9, no. 6, June 2017.
- IEEE P7005™, Standards Project 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.

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 the global community to recognize which sets of values guide the design, and whether or not A/IS will generate problematic, i.e., 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, such as utilitarianism, deontology, and virtue ethics. There is also a need to include other traditions of ethics in RI, such as those inherent to Buddhism, Confucianism, and Ubuntu traditions.

However, this venture poses problematic assumptions even before the issue above can be explored. In classifying Western values, we group together thousands of years of independent and disparate ideas originating from the Greco-Roman philosophical tradition with their Christian-infused cultural heritage and then the break from that heritage with the Enlightenment. What is it that one refers to by the term "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 "Eastern" ethical traditions. Suggest instead: If one is referring to Western values, one must designate which values and to whom they belong. Additionally, there is a danger in the field of intercultural information ethics, however

Classical Ethics in A/IS

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 importance to first properly distinguish between subtleties in Western ethics as a discipline and morality as its object or subject matter. It is also important to differentiate between philosophical or scientific ethics and theological ethics. As noted above, the relationship between assumed moral customs, the ethical critique of those customs, and the law is an established methodology in scientific communities. Western and Eastern philosophy are very different, just like Western and Eastern ethics. Western philosophical ethics use scientific methods such as the logical, discursive, and dialectical approach (models of normative ethics) alongside the analytical and hermeneutical approaches. 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 A/IS 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 value monopolies to find alternative methods of accommodating different value systems. Liberal values arose out of conflicts

Classical Ethics in A/IS

of cultural and subcultural differences and are designed to be accommodating enough to include a rather wide range of differences.

RI enables policymakers, 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 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.

In addition, many non-Western ethics traditions, including the Buddhist and Ubuntu traditions highlighted below, view "relationship" as a foundationally important concept to ethical discourse. One of the key parameters of intercultural information ethics and RI research must be to identify main commonalities of "relationship" approaches from different cultures and how to operationalize them for A/IS to complement classical methodologies of deontological and teleological ethics. Different cultural perceptions of time may influence "relationship" approaches and impact how A/IS are perceived and integrated, e.g., technology as part of linear progress in the West; inter-generational needs and principles of respect and benevolence in Chinese culture determining current and future use of technology.

Recommendations

In order to enable a cross-cultural dialogue of ethics in technology, discussions on ethics and A/IS must first return to normative foundations of RI to address the notion of "responsible innovation" from a range of value systems not predominant in Western classical ethics. Together with acknowledging differences, a special focus on commonalities in the intercultural understanding of the concept of "relationship" must complement the process.

Further Resources

- J. Bielby, "Comparative Philosophies in Intercultural Information Ethics," Confluence: Journal of World Philosophies, vol. 2, 2016.
- W. B. Carlin and K. C. Strong, "A Critique of Western Philosophical Ethics: Multidisciplinary Alternatives for Framing Ethical Dilemmas," Journal of Business Ethics, vol. 14, no. 5, pp. 387-396, May 1995.
- C. Ess, "Lost in translation"?: Intercultural dialogues on privacy and information ethics (introduction to special issue on privacy and data privacy protection in Asia)," Ethics and Information Technology, vol. 7, no. 1, pp. 1-6, March 2005.
- S. Hongladarom, "Intercultural Information Ethics: A Pragmatic Consideration," in Information Cultures in the Digital Age. Wiesbaden, Germany: Springer Fachmedien, 2016, pp. 191-206.
- L. G. Rodríguez and M. Á. P. Álvarez, Ética Multicultural y Sociedad en Red. Madrid: Fundación Telefónica, 2014.

Classical Ethics in A/IS

- P. H. Wong, "What Should We Share?: Understanding the Aim of Intercultural Information Ethics," ACM SIGCAS Computers and Society, vol. 39, no. 3 pp. 50-58, Dec. 2009.
- S. A. Wilson, "Conformity, Individuality, and the Nature of Virtue: A Classical Confucian Contribution to Contemporary Ethical Reflection," The Journal of Religious Ethics, vol. 23, no. 2, pp. 263-289, 1995.
- P. H. Wong, "Responsible Innovation for Decent Nonliberal Peoples: A Dilemma?" Journal of Responsible Innovation, vol. 3, no. 2, pp. 154-168, July 2016.
- R. B. Zeuschner, Classical Ethics, East and West: Ethics from a Comparative Perspective. Boston, MA: McGraw-Hill, 2000.
- S. Mattingly-Jordan, "Becoming a Leader in Global Ethics," IEEE, 2017.

Issue: The Application of Classical Buddhist Ethical Traditions to A/IS Design

Background

According to Buddhism, the field of 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 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, that ensnare one's self against freedom. In order to attain liberation, one engages oneself in mindful behavior (ethics), concentration (meditation), and what is deemed in Buddhism as "wisdom", a term that remains ambiguous in Western scientific approaches to ethics.

Thus ethics in Buddhism are concerned exclusively with how to attain the goal of liberation, or freedom. In contrast to Western ethics, Buddhist ethics are not concerned with theoretical questions on 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 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 that it be considered in a scientific

Classical Ethics in A/IS

context. Such assumptions put it at odds 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 contain potential for conflict with Western ethical value systems which are founded on ideas of questioning moral and epistemological assumptions. Buddhist ethics are 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 the 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 from the concept of privacy as often 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 of certain values that are found to be necessary for a well-functioning society to 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. Such a state is the 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 ataraxia, i.e., 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 as 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.

Classical Ethics in A/IS

Recommendations

In considering the nature of interactions between human and autonomous systems, 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. It is recommended as an additional methodology, along with Western-value methodologies, to address human/computer interactions.

Further Resources

- R. Capurro, "Intercultural Information Ethics: Foundations and Applications," Journal of Information, Communication & Ethics in Society, vol. 6, no. 2, pp. 116-126, 2008.
- C. Ess, "Ethical Pluralism and Global Information Ethics," Ethics and Information Technology, vol. 8, no. 4, pp. 215-226, Nov. 2006.
- S. Hongladarom, "Intercultural Information Ethics: A Pragmatic Consideration," in Information Cultures in the Digital Age, K. M. Bielby, Ed. Wiesbaden, Germany: Springer Fachmedien Wiesbaden, 2016, pp. 191-206.
- S. Hongladarom, J. Britz, "Intercultural Information Ethics," International Review of Information Ethics, vol. 13, pp. 2-5, Oct. 2010.
- M. Nakada, "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, pp. 300-314, 2010.
- M. Mori, 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). In applying that question to A/IS, it reads: "If an ethic focused on the good of community, interpreted philosophically as a moral theory, is applied to A/IS, what would the implications be on the duties of various agents?" Agents, in this regard, would therefore be the following:

- Members of the A/IS research community
- A/IS programmers/computer scientists
- A/IS end-users
- A/IS themselves

Classical Ethics in A/IS

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. It is a commonly held maxim in the Ubuntu tradition that, "to be is to belong to a community and participate." As the saying goes, *motho ke motho ka batho babang,* or, "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. Among 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

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. Ubuntu, as collective human dignity, may offer a way of understanding the impact of A/IS on humankind, e.g., the need for human moral and legal agency; human life and death decisions to be taken by humans rather than A/IS.

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). As Tutu notes, "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, they will have to adhere to the principles of community, identity, and solidarity with others. On the other hand, they will also need to be cognizant of, and sensitive toward, the potential for community-based ethics to exclude individuals on the basis that they do not belong or fail to meet communitarian standards. For example,

Classical Ethics in A/IS

would this mean the excluded individual lacks personhood and as a consequence would not be able to benefit from community-based A/IS initiatives? How would community-based A/IS programming avoid such biases against individuals?

While virtue ethics question the goal or purpose of A/IS and deontological ethics question 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, they will be detrimental to the identity of this community when it comes to job losses, poverty, lacks in education, and lacks in skills training. However, should A/IS seek to supplement the community by means of ease of access, support systems, and more, then it cannot be argued that they will be detrimental. In between these two motivators is a safeguarding issue about how to avoid excluding individuals from accessing community-based A/IS initiatives. It therefore becomes imperative that whoever designs the systems must work closely both with ethicists and the target community, audience, or end-user to ascertain whether their needs are identified and met.

Recommendations

It is recommended that a concerted effort be made toward the study and publication of literature addressing potential relationships between Ubuntu and other instances of African 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 objectives, products, and services are aligned with the needs of the end-users and target communities.

Further Resources

- D. W. Lutz, "African Ubuntu Philosophy and Global Management," Journal of Business Ethics, vol. 84, pp. 313-328, Oct. 2009.
- T. Metz, "African Ethics and Journalism Ethics: News and Opinion in Light of Ubuntu," Journal of Media Ethics: Exploring Questions of Media Morality, vol. 30 no. 2, pp. 74-90, April 2015.
- T. Metz, "Ubuntu as a moral theory and human rights in South Africa," African Human Rights Law Journal, vol. 11, no. 2, pp. 532-559, 2011.
- R. Nicolson, Persons in Community: African Ethics in a Global Culture. Scottsville, South Africa: University of KwaZulu-Natal Press, 2008.
- A. Shutte, Ubuntu: An Ethic for a New South Africa. Dorpspruit, South Africa: Cluster Publications, 2001.
- D. Tutu, No Future without Forgiveness. London: Rider, 1999.
- O. Ulgen, "Human Dignity in an Age of Autonomous Weapons: Are We in Danger of Losing an 'Elementary Consideration of Humanity'?" in How International Law Works in Times of Crisis, I. Ziemele and G. Ulrich, Eds. Oxford: Oxford University Press, 2018, pp. 242-272.

Classical Ethics in A/IS

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 catalyst for Japanese robot and autonomous systems culture, a culture that naturally stems from the traditional Japanese concept of *karakuri ningyo* (automata). Popular Japanese artificial intelligence, 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 arguably 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 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—the first, to its Greek etymology that combines andras ("ἀνδρας"), or man, and gynoids/gyni ("γυνή"), or woman; the second, via automatons and toys as per U.S. patent developers in the 1800s; and the third 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 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 (robots). Artifacts are as much a part of nature in Shinto as animals, and they are considered naturally beautiful rather than falsely artificial.

A potential conflict between Western and Japanese concepts of nature and artifact arises when the two traditions are compared

Classical Ethics in A/IS

and contrasted, especially in the exploration of artificial intelligence. While in Shinto, the artifact as "artificial" represents creation and authentic being, with implications for defining autonomy, the same artifact is designated as secondary and often 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".

Recommendations

Where Japanese culture leads the way in the synthesis of traditional value systems and technology, we recommend that people involved with 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 A/IS.

Further Resources

- R. M. Geraci, "Spiritual Robots: Religion and Our Scientific View of the Natural World," Theology and Science, vol. 4, no. 3, pp. 229-246, 2006.
- D. F. Holland-Minkley, "God in the Machine: Perceptions and Portrayals of Mechanical Kami in Japanese Anime." Ph.D. dissertation, University of Pittsburgh, Pittsburgh, PA, 2010.
- C. B. Jensen and A. Blok, "Techno-Animism in Japan: Shinto Cosmograms, Actor-Network Theory, and the Enabling Powers of Non-Human Agencies," Theory, Culture & Society, vol. 30, no. 2, pp. 84-115, March 2013.
- F. Kaplan, "Who Is Afraid of the Humanoid? Investigating Cultural Differences in the Acceptance of Robots," International Journal of Humanoid Robotics, vol. 1, no. 3, pp. 465-480, 2004.
- S. G. Tzafestas, Roboethics: A Navigating Overview. Cham, Switzerland: Springer, 2015.
- G. Veruggio and K. Abney, "22 Roboethics: The Applied Ethics for a New Science," in Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge, MA: MIT Press, 2011, p. 347.

Section 3—Classical Ethics for a Technical World

Issue: Maintaining Human Autonomy

Background

A/IS 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 functions. 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 A/IS 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 A/IS 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 bioinformation 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 bioinformation in targeted advertising campaigns that seek to influence the user in real-time feedback loops based on the user's biological reactions such as

Classical Ethics in A/IS

pupil dilation, body temperature, and 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. Should there be a precedent for the A/IS user to know when they are interacting with an algorithm? What about consent?

The responsibility for the behavior of algorithms remains with the designer, 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 others? Who decides on the parameters? Is it the machine itself? Such questions become key factors in conversations around ethical standards.

Recommendations

A two-step process is recommended to maintain human autonomy in A/IS. The creation of an ethics-by-design methodology is the first step to addressing human autonomy in A/IS, 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 the creation of a pointed and widely applied education curriculum that spans grade school 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

- B. van den Berg and J. de Mul, "Remote Control. Human Autonomy in the Age of Computer-Mediated Agency," in Law, Human Agency and Autonomic Computing: The Philosophy of Law Meets the Philosophy of Technology, M. Hildebrandt and A. Rouvroy, Eds. London: Routledge, 2011, pp. 46-63.
- L. Costa, "A World of Ambient Intelligence," in Virtuality and Capabilities in a World of Ambient Intelligence. Cham, Switzerland: Springer International, 2016, pp. 15-41.
- P. P. Verbeek, "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, M. Hildebrandt and A. Rouvroy, Eds. New York: Routledge, 2011.

Classical Ethics in A/IS

- D. Reisman, J. Schultz, K. Crawford, and M. Whittaker, "Algorithmic Impact Assessments: A practical Framework for Public Agency Accountability," AI NOW, April 2018.
- A. Chaudhuri, "Philosophical Dimensions of Information and Ethics in the Internet of Things (IoT) Technology," EDPACS, vol. 56, no. 4, pp. 7-18, Nov. 2017.

Issue: Implications of Cultural Migration in A/IS

Background

In addition to developing an understanding of A/IS via different cultures, it is crucial to understand how A/IS are shaped and reshaped —how they affect and are affected by—human mobility and cultural diversity through active immigration. The effect of human mobility on state systems reliant on A/IS impacts the State structure itself, and thus the systems that the structure relies on, in the end influencing everything from democracy to citizenship. Where the State, through A/IS, invests in and gathers big data through mechanisms for registration and identification of people, mainly immigrants, human mobility becomes a foundational component in a system geared toward the preservation of human dignity.

Traditional national concerns reflect two information foundations: information produced for human rights and information produced for national sovereignty. In the second foundation, State borders are considered the limits from which political governance is defined in terms of security. The preservation of national sovereignty depends on the production and domination of knowledge. In the realm of migratory policies, knowledge is created to measure people in transit: collecting, treating, and transferring information about territory and society.

Knowledge organization has been the paramount pillar of scientific thought and scientific practice since the beginning of written civilization. Any scientific and technological development has only been possible through information policies that include the establishment of management processes to systematize them, and the codification of language. For the Greeks, this process was closely associated with the concept of *arete*, or the excellence of one's self in politics as congregated in the *polis*. The notion of *polis* is as relevant as ever in the digital age with the development of digital technologies and the discussions around morality in A/IS. Where the systematization of knowledge is potentially freely created, the advent of the Internet and its flows are difficult to control. Ethical issues about the production of information are becoming paramount to our digital society.

The advancement of the fields of science and technology has not been followed by innovations in the political community, and the technical community has repeatedly tabled academic discussions about the hegemony of technocracy over policy issues, restricting the space of the policy arena and valorizing excessively technic solutions for human problems. This monopoly alters conceptions of morality, relocating the locus of the Kantian "Categorical Imperative", causing the tension among different social and political contexts to become more pervasive.

Classical Ethics in A/IS

Current global migration dynamics have been met by unfavorable public opinion based in ideas of crisis and emergency, a response vastly disproportionate to what statistics have shown to be the reality. In response to these views, A/IS are currently designed and applied to measure, calculate, identify, register, systematize, normalize, and frame both human rights and security policies. This is largely no different of a process than what has been practiced since the period of colonialism. It includes the creation and implementation of a set of ancient and new technologies. Throughout history, mechanisms have been created firstly to identify and select individuals who share certain biological heritage, and secondly to individuals and social groups, including biological characteristics.

Information is only possible when materialized as an infrastructure supported by ideas in action as a "communicative act", which Habermas (1968) identifies in Hegel's work, converging three elements in human-in-the-world relationships: symbol, language, and labor. Information policies reveal the importance and the strength in which technologies influence economic, social, cultural, identity, and ethnic interactions.

Traditional mechanisms used to control migration, such as the passport, are associated with globally established walls and fences. The more intense human mobility becomes, the more amplified are the discourses to discourage it, restricting human migrations, and deepening the need for an ethics related to conditions of citizenship. Together with the building of walls, other remote technologies are developed to monitor and surveil borders, buildings, and streets, also impacting ideas and moral presumptions of citizenship. Closed Circuit Television(CCTV), Unmanned Aerial Vehicles (UAVs), and satellites allow data transference in real time to databases, cementing the backbone that A/IS draws from, often with bias as per the expectations of developed countries. This centrality of data sources for A/IS expresses a divide between developed and underdeveloped countries, particularly as relevant to the refugee.

Information is something that links languages, habits, customs, identification, and registration technologies. It provokes a reshaping of the immigrants' and refugees' citizenship and their value as people in terms of their citizenship, as they seek forms of surviving in, and against, the restrictions imposed by A/IS for surveillance and monitoring in an enlarged and more complex cosmopolis.

An understanding of the impact of A/IS on migration and mobile populations, as used in state systems, is a critical first step to consider if systems are to become truly autonomous and intelligent, especially beyond the guidance of human deliberation. Digital technology systems used to register and identify human mobility, including refugees and other displaced populations, are not autonomous in the intelligent sense, and are dependent on the biases of worldviews around immigration. In this aspect, language is the locus where this dichotomy has to be considered to understand the diversity of morals when there are contacts among different cultures.

Classical Ethics in A/IS

Recommendations

Is it recommended that the State become a proactive player in the globalized processes of A/IS for migrant and mobile populations, introducing a series of mechanisms that limit the segregation of social spaces and groups, and consider the biases inherent in surveillance for control.

Further Resources

- I. About and V. Denis, Histoire de l'identification des personnes. Paris: La Découverte, 2010.
- I. About, J. Brown, G. Lonergan, Identification and Registration Practices in Transnational Perspective: People, Papers and Practices. London: Palgrave Macmillan, 2013, pp. 1-13.
- D. Bigo, "Security and Immigration: Toward a Critique of the Governmentality of Unease," in Alternatives, Special Issue, no. 27. pp. 63-92, 2002.
- R. Capurro, "Citizenship in the Digital Age," in Information Ethics, Globalization and Citizenship, T. Samek and L. Schultz, Eds. Jefferson NC: McFarland, 2017, pp. 11-30.
- R. Capurro, "Intercultural Information Ethics," in Localizing the Internet: Ethical Aspects in Intercultural Perspective, R. Capurro, J. Frühbauer, and T. Hausmanninger, Eds. Munich: Fink, 2007, pp. 21-38.
- UN High Commissioner for Refugees (UNHCR), Policy on the Protection of Personal Data of Persons of Concern to UNHCR, May 2015.

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 internally 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 approach that does not rely on a decision

Classical Ethics in A/IS

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.

Recommendations

Program autonomous systems to be able to recognize user behavior for the purposes of predictability, traceability, and accountability 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

- M. A. Boden, Ed. The Philosophy of Artificial Life. Oxford, U.K.: Oxford University Press, 1996.
- C. Castelfranchi, "Modelling Social Action for AI Agents," Artificial Intelligence, vol. 103, no.1-2, pp. 157-182, 1998.
- W. D. Christensen and C. A. Hooker, "Anticipation in Autonomous Systems: Foundations for a Theory of Embodied Agents," International Journal of Computing Anticipatory Systems, vol. 5, pp. 135-154, Dec. 2000.
- K. G. Coleman, "Android Arete: Toward a Virtue Ethic for Computational Agents," Ethics and Information Technology, vol. 3, no. 4, pp. 247-265, 2001.
- J. G. Lennox, "Aristotle on the Biological Roots of Virtue," Biology and the Foundations of Ethics, J. Maienschein and M. Ruse, Eds. Cambridge, U.K.: Cambridge University Press, 1999, pp. 405-438.
- L. Muehlhauser and L. Helm, "The Singularity and Machine Ethics," in Singularity Hypotheses, A. H. Eden, J. H. Moor, J. H. Soraker, and E. Steinhart, Eds. Berlin: Springer, 2012, pp. 101-126.
- D. Vernon, G. Metta, and G. Sandini, "A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents," IEEE Transactions on Evolutionary Computation, vol. 11, no. 2, pp. 151-180, April 2007.

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 metarules. For example, a rule might take the form "Don't lie!", whereas a metarule 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."

Classical Ethics in A/IS

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 is 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. 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.

Recommendations

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 determinable formal systems.

Further Resources

- C. Allen, I. Smit, and W. Wallach, "Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches," Ethics and Information Technology, vol. 7, no. 3, pp. 149-155, 2005.
- O. Bendel, Die Moral in der Maschine: Beiträge zu Roboter-und Maschinenethik. Heise Medien, 2016.
- O. Bendel, Oliver, Handbuch Maschinenethik. Wiesbaden, Germany: Springer VS, 2018.
- M. Fisher, L. Dennis, and M. Webster, "Verifying Autonomous Systems," Communications of the ACM, vol. 56, no. 9, pp. 84-93, Sept. 2013.
- B. M. McLaren, "Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions," IEEE Intelligent Systems, vol. 21, no. 4, pp. 29-37, July 2006.
- M. A. Perez Alvarez, "Tecnologías de la Mente y Exocerebro o las Mediaciones del Aprendizaje," 2015.
- E. L. Rissland and D. B. Skalak, "Combining Case-Based and Rule-Based Reasoning: A Heuristic Approach." Proceedings of the 11th International Joint Conference on Artificial Intelligence, IJCAI 1989, Detroit, MI, August 20-25, 1989, San Francisco, CA: Morgan Kaufmann Publishers Inc., 1989. pp. 524-530.

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Classical Ethics in A/IS Committee

- **Jared Bielby** (Chair) – President, Netizen Consulting Ltd; Chair, International Center for Information Ethics; editor, *Information Cultures in the Digital Age*
- **Soraj Hongladarom** (Co-chair) – President at The Philosophy and Religion Society of Thailand
- **Miguel Á. Pérez Álvarez** – Professor of Technology in Education, Colegio de Pedagogía, Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México
- **Oliver Bendel** – Professor of Information Systems, Information Ethics and Machine Ethics, University of Applied Sciences and Arts Northwestern Switzerland FHNW
- **Dr. John T. F. Burgess** – Assistant Professor / Coordinator for Distance Education, School of Library and Information Studies, The University of Alabama
- **Rafael Capurro** – Founder, International Center for Information Ethics
- **Corinne Cath-Speth** – PhD student at Oxford Internet Institute, The University of Oxford, Doctoral student at the Alan Turing Institute, Digital Consultant at ARTICLE 19
- **Dr. Paola Di Maio** – Center for Technology Ethics, ISTCS.org UK and NCKU Taiwan
- **Robert Donaldson** – Independent Computer Scientist, BMRILLC, Hershey, PA

Classical Ethics in A/IS

- **Rachel Fischer** – Research Officer: African Centre of Excellence for Information Ethics, Information Science Department, University of Pretoria, South Africa.
- **Dr. D. Michael Franklin** – Assistant Professor, Kennesaw State University, Marietta Campus, Marietta, GA
- **Wolfgang Hofkirchner** – Associate Professor, Institute for Design and Technology Assessment, Vienna University of Technology
- **Dr. Tae Wan Kim** – Associate Professor of Business Ethics, Tepper School of Business Carnegie Mellon University
- **Kai Kimppa** – University Research Fellow, Information Systems, Turku School of Economics, University of Turku
- **Sara R. Mattingly-Jordan** – Assistant Professor Center for Public Administration & Policy, Virginia Tech
- **Dr Neil McBride** – Reader in IT Management, School of Computer Science and Informatics, Centre for Computing and Social Responsibility, De Montfort University
- **Bruno Macedo Nathansohn** – Perspectivas Filosóficas em Informação (Perfil-i); Brazilian Institute of Information in Science and Technology (IBICT)
- **Marie-Therese Png** – PhD Student, Oxford Internet Institute, PhD Intern, DeepMind Ethics & Society
- **Derek Poitras** – Independent Consultant, Object Oriented Software Development
- **Samuel T. Segun** – PhD Candidate, Department of Philosophy, University of Johannesburg. Fellow, Philosophy Node of the Centre for Artificial Intelligence Research (CAIR) at the University of Pretoria and Research fellow at the Conversational School of Philosophy (CSP)
- **Dr. Ozlem Ulgen** – Reader in International Law and Ethics, School of Law, Birmingham City University
- **Kristene Unsworth** – Assistant Professor, The College of Computing & Informatics, Drexel University
- **Dr. Xiaowei Wang** – Associate professor of Philosophy, Renmin University of China
- **Dr Sara Wilford** – Senior Lecturer, Research Fellow, School of Computer Science and Informatics, Centre for Computing and Social Responsibility, De Montfort University
- **Pak-Hang Wong** – Research Associate, Department of Informatics, University of Hamburg
- **Bendert Zevenbergen** – Oxford Internet Institute, University of Oxford & Center for Information Technology Policy, Princeton University

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared.*

Endnotes

[1] This edition of "Classical Ethics in A/IS" does not (and could not) aspire to universal coverage of all of the world's traditions in the space available to us. Future editions will touch on several other traditions, including Judaism and Islam.

[2] R. Von Schomberg, "Prospects for Technology Assessment in a Framework of Responsible Research and Innovation" in Technikfolgen Abschätzen Lehren: Bildungspotenziale Transdisziplinärer Methode. Wiesbaden, Germany: Springer VS, 2011, pp. 39-61.

Well-being

Prioritizing ethical and responsible artificial intelligence has become a widespread goal for society. Important issues of transparency, accountability, algorithmic bias, and value systems 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 policymakers alike is how to assess, understand, measure, monitor, safeguard, and improve the well-being impacts of A/IS on humans. Finding the answer to this question is further complicated when A/IS are within a holistic and interconnected framework of well-being in which individual well-being is inseparable from societal, economic, and environmental systems.

For A/IS to demonstrably advance well-being, we need consistent and multidimensional indicators that are easily implementable by the developers, engineers, and designers who are building our future. This chapter is intended for such developers, engineers, and designers—referred to in this chapter as "A/IS creators". Those affected by A/IS are referred to as "A/IS stakeholders".

A/IS technologies affect human agency, identity, emotion, and ecological systems in new and profound ways. Traditional metrics of success are not equipped to ensure A/IS creators can avoid unintended consequences or benefit from unexpected innovation in the algorithmic age. A/IS creators need expanded ways to evaluate the impact of their products, services, or systems on human well-being. These evaluations must also be done with an understanding that human well-being is deeply linked to the well-being of society, economies, and ecosystems.

Today, A/IS creators largely measure success using metrics including profit, gross domestic product (GDP), consumption levels, and occupational safety. While important, these metrics fail to encompass the full spectrum of well-being impacts on individuals and society, such as psychological, social, and environmental factors. Where the priority given to these factors is not equal to that given to fiscal metrics of success, A/IS creators risk causing or contributing to negative and irreversible harms to our people and our planet.

When A/IS creators are not aware that well-being indicators, in addition to traditional metrics, can provide guidance for their work, they are also missing out on innovation that can increase well-being and societal value. For instance, while it is commonly recognized that autonomous vehicles will save lives when safely deployed, a topic of less frequent discussion is how self-

Well-being

driving cars also have the potential to help the environment by reducing greenhouse gas emissions and increasing green space. Autonomous vehicles can also positively impact well-being by increasing work-life balance and enhancing the quality of time spent during commutes.

Unless A/IS creators are made aware of the existence of alternative measures of progress, the value they provide, and the way they can be incorporated into A/IS work, technology and society will continue to rely upon traditional metrics of success. In an era where innovation is defined by holistic prosperity, alternative measures are needed more now than ever before. The 2009 Report by the Commission on the Measurement of Economic Performance and Social Progress which contributed substantially to the worldwide movement of governments using wider measures of well-being, states, "What we measure affects what we do; and if our measurements are flawed, decisions may be distorted."

We believe that A/IS creators can profoundly increase human and environmental flourishing by prioritizing well-being metrics as an outcome in all A/IS system designs—now and for the future. *The primary intended audience for this chapter is A/IS creators who are unfamiliar with the term "well-being" as it is used in the field of positive psychology and well-being studies. Our initial goal is to provide a broad introduction to qualitative and quantitative metrics and applications of well-being to educate and inspire A/IS creators. We do not prioritize or advocate for any specific indicator or methodology. For further elaboration on the definition of well-being, please see the first Issue listed in Section 1.*

This chapter is divided into two main sections:

- The Value of Well-being Metrics for A/IS Creators
- Implementing Well-being Metrics for A/IS Creators

The following resources are available online to provide readers with an introduction to existing well-being metrics and tools currently in use:

- The State of Well-being Metrics
- The Happiness Screening Tool for Business Product Decisions
- Additional Resources: Standards Development Models and Frameworks

Section 1—The Value of Well-being Metrics for A/IS Creators

Well-being metrics provide a broader perspective for A/IS creators than they normally might be familiar with in evaluating their products. This broader perspective unlocks greater opportunities to assure a positive impact of A/IS on human well-being, while minimizing the risk of unintended negative outcomes. This section defines well-being, discusses the value of well-being metrics to A/IS creators, and notes how similar frameworks like sustainability and human rights can be complemented by incorporating well-being metrics.

Definition of Well-being

For the purposes of *Ethically Aligned Design*, the term "well-being" refers to an evaluation of the general quality of life of an individual and the state of external circumstances. The conception of well-being encompasses the full spectrum of personal, social, and environmental factors that enhance human life and on which human life depend. The concept of well-being shall be considered distinct from moral or legal evaluation.

Issue: There is ample and robust science behind well-being metrics and their use by international and national institutions. However, A/IS creators are often unaware that well-being metrics exist, or that they can be used to plan, develop, and evaluate technology.

Background

The concept of well-being refers to an evaluation of the general goodness of the state of an individual or community and is distinct from moral or legal evaluation. A well-being evaluation takes into account major aspects of a person's life, such as their happiness, success in their goals, and their overall positive functioning in their environment. There is now a thriving area of scientific research into the psychological, social, behavioral, economic, and environmental determinants of human well-being.

Well-being

The term "well-being" is defined and used in various ways across different contexts and fields. For example: economists identifying economic welfare with levels of consumption and economic vitality, psychologists highlighting subjective experience, and sociologists emphasizing living, labor, political, social, and environmental conditions. We do not take a stand on any specific measure of well-being. The metrics listed below are an incomplete list and provided as a starting point for further inquiry. Among these are subjective well-being indicators, measures of quality of life, social progress and capabilities, and many more.

There is now sufficient consensus among scientists that well-being can be reliably measured. Well-being measures differ in the number and the intricacy of indicators they employ. Short questionnaires of life satisfaction have emerged as particularly popular, although they do not reflect all aspects of well-being. While recognizing a scope for differences across well-being indicators, we note that the richest conception of well-being encompasses the full spectrum of personal, social, and environmental goods that enhance human life.

We encourage A/IS creators to consider the wide range of available indicators and select those most relevant and revealing for particular stages of the A/IS technology's life cycle and the particular context for the technology's use and evaluation. That is, measures of well-being that may be well-suited to wealthy, industrialized nations may be less applicable in low- and middle-income countries, and vice versa.

Among the most important and recognized aspects of well-being are (in alphabetical order):

- Community: Belonging, Crime & Safety, Discrimination & Inclusion, Participation, Social Support
- Culture: Identity, Values
- Economy: Economic Policy, Equality & Environment, Innovation, Jobs, Sustainable Natural Resources & Consumption & Production, Standard of Living
- Education: Formal Education, Lifelong Learning, Teacher Training
- Environment: Air, Biodiversity, Climate Change, Soil, Water
- Government: Confidence, Engagement, Human Rights, Institutions
- Human Settlements: Energy, Food, Housing, Information & Communication Technology, Transportation
- Physical Health: Health Status, Risk Factors, Service Coverage
- Psychological Health: Affect (feelings), Flourishing, Mental Illness & Health, Satisfaction with Life
- Work: Governance, Time Balance, Workplace Environment

Well-being

In an effort to provide a basic orientation to well-being metrics, information about well-being indicators can be segmented into four categories:

1. Subjective or survey-based indicators
Survey-based well-being indicators, subjective well-being (SWB) indicators, and multidimensional measurements of aspects of well-being, are being used by national institutions, international institutions, and governments to better understand levels of psychological well-being within countries and aspects of a country's population. These indicators are also being used to understand people's satisfaction in specific domains of life. Examples of surveys that include survey-based well-being indicators and SWB indicators include the European Social Survey, Bhutan's Gross National Happiness Indicators, well-being surveys created by The UK Office for National Statistics, and many more.

Survey-based metrics are also employed in the field of positive psychology and in the World Happiness Report. The data are employed by researchers to understand the causes, consequences, and correlates of well-being. Data gathered from surveys tend to address concerns, such as day-to-day experience, overall satisfaction with life, and perceived flourishing. The findings of these researchers provide crucial and necessary guidance because they often diverge from and complement the understanding of traditional conditions, such as economic growth.

2. Objective indicators
Objective indicators of quality of life have typically incorporated areas such as income, consumption, health, education, crime, housing, etc. These indicators have been used to understand conditions that support the well-being of countries and populations, and to measure the societal and environmental impact of companies. They are in use by organizations like the OECD with their Better Life Index, which also includes survey-based well-being indicators and SWB indicators, and the United Nations with their Sustainable Development Goals Indicators (formerly the Millennium Development Goals). 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 reflecting both objective aspects of quality of life and people's subjective evaluation of these. 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. Some subjective and objective indicators are also composite indicators, such as Bhutan's Gross National Happiness Index and the OECD's Better Life Index.

4. Social media sourced data
Social media can be used to measure the well-being of a geographic region or demographic group, based on sentiment analysis of publicly available data. Examples include the Hedonometer and the World Well-being Project.

Well-being

Recommendation

A/IS creators should prioritize learning about well-being concepts, scientific learnings, research findings, and well-being metrics as potential determinants for how they create, deploy, market, and monitor their technologies, and ensuring their stakeholders learn the same. This process can be expedited if Standards Development Organizations (SDOs), such as the IEEE Standards Association, or other institutions such as the Global Reporting Initiative (GRI) or B-Corp, create certifications, guidelines, and standards that for the use of holistic, well-being metrics for A/IS in the public and private sectors.

Further Resources

- The IEEE P7010™ Standards Project for Well-being Metric for Autonomous/Intelligent Systems, was 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)*

Issue: Increased awareness and application of well-being metrics by A/IS creators can create greater value, safety, and relevance to corporate communities and other organizations in the algorithmic age.

Background

While many organizations in the private and public sectors are increasingly aware of the need to incorporate well-being measures as part of their efforts, the reality is that bottom line, quarterly-driven shareholder growth remains a dominant goal and metric. Short term growth is often the priority in the private sector and public sector. As long as organizations exist in a larger societal system which prioritizes financial success, these companies will remain under pressure to deliver financial results that do not fully incorporate societal and environmental impacts, measurements, or priorities.

Rather than focus solely on the negative aspects of how A/IS could harm humans and environments, we seek to explore how the implementation of well-being metrics can help A/IS to have a measurable, positive impact on human well-being as well as on systems and organizations. Incorporation of well-being goals and measures beyond what is strictly required can benefit both private sector organizations' brands and public sector organizations' stability and reputation, as well as help realize financial

Well-being

savings, innovation, trust, and many other benefits. For instance, a companion robot outfitted to support seniors in assisted living situations might traditionally be launched with a technology development model that was popularized by Silicon Valley known as "move fast and break things". The A/IS creator who rushed to bring the robot to market faster than the competition and who was unaware of well-being metrics, may have overlooked critical needs of the seniors. The robot might actually hurt the senior instead of helping by exacerbating isolation or feelings of loneliness and helplessness. While this is a hypothetical scenario, it is intended to demonstrate the value of linking A/IS design to well-being indicators.

By prioritizing largely fiscal metrics of success, A/IS devices might fail in the market because of limited adoption and subpar reception. However, if during use of the A/IS product, success were measured in terms of relevant aspects of well-being, developers and researchers could be in a better position to attain funding and public support. Depending on the intended use of the A/IS product, well-being measures that could be used extend to emotional levels of calm or stress; psychological states of thriving or depression; behavioral patterns of engagement in community or isolation; eating, exercise and consumption habits; and many other aspects of human well-being. The A/IS product could significantly improve quality of life guided by metrics from trusted sources, such as the World Health Organization, European Social Survey, and Sustainable Development Goal Indicators.

Thought leaders in the corporate arena have recognized the multifaceted need to utilize metrics beyond fiscal indicators. PricewaterhouseCoopers defines "total impact" as a "holistic view of social, environmental, fiscal and economic dimensions—the big picture". Other thought-leading organizations in the public sector, such as the OECD, demonstrate the desire for business leaders to incorporate metrics of success beyond fiscal indicators for their efforts, exemplified in their 2017 workshop, Measuring Business Impacts on People's Well-Being. The B-Corporation movement has 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, B-Corps are defining their brands by provably aligning their efforts with wider measures of well-being.

Recommendations

A/IS creators should work to better understand and apply well-being metrics in the algorithmic age. Specifically:

- A/IS creators should work directly with experts, researchers, and practitioners in well-being concepts and metrics to identify existing metrics and combinations of indicators that would bring support a "triple bottom line", i.e., accounting for economic, social, and environmental impacts, approach to well-being. However, well-being metrics should only be used with consent, respect for privacy, and with strict standards for collection and use of these data.
- For A/IS to promote human well-being, the well-being metrics should be chosen in collaboration with the populations most affected by those systems—the A/IS

Well-being

stakeholders—including both the intended end-users or beneficiaries and those groups whose lives might be unintentionally transformed by them. This selection process should be iterative and through a learning and continually improving process. In addition, "metrics of well-being" should be treated as vehicles for learning and potential mid-course corrections. The effects of A/IS on human well-being should be monitored continuously throughout their life cycles, by A/IS creators and stakeholders, and both A/IS creators and stakeholders should be prepared to significantly modify, or even roll back, technology that is shown to reduce well-being, as defined by affected populations.

- A/IS creators in the business or academic, engineering, or policy arenas are advised to review the additional resources on standards development models and frameworks at the end of this chapter to familiarize themselves with existing indicators relevant to their work.

Further Resources

- PricewaterhouseCoopers (PwC). Managing and Measuring Total Impact: A New Language for Business Decisions, 2017.
- World Economic Forum. The Inclusive Growth and Development Report 2017, Geneva, Switzerland: World Economic Forum, January 16, 2017.
- OECD Guidelines on Measuring Subjective Well-being, 2013.
- National Research Council. Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience. DC: The National Academies Press, 2013.

Issue: A/IS creators have opportunities to safeguard human well-being by ensuring that A/IS does no harm to earth's natural systems or that A/IS contributes to realizing sustainable stewardship, preservation, and/or restoration of earth's natural systems. A/IS creators have opportunities to prevent A/IS from contributing to the degradation of earth's natural systems and hence losses to human well-being.

Background

It is unwise, and in truth impossible, to separate the well-being of the natural environment of the planet from the well-being of humanity. A range of studies, from the historic to more recent, prove that ecological collapse endangers human existence. Hence, the concept of well-being should encompass planetary well-being. Moreover, biodiversity and ecological integrity have intrinsic merit beyond simply their instrumental value to humans.

Technology has a long history of contributing to ecological degradation through its role in expanding the scale of resource extraction and environmental pollution, for example, the immense power needs of network computing, which leads to climate change, water scarcity, soil degradation, species extinction, deforestation,

Well-being

biodiversity loss, and destruction of ecosystems which in turn threatens humankind in the long run. These and other costs are often considered externalities and often do not figure into decisions or plans. At the same time, there are many examples, such as photovoltaics and smart grid technology that present potential ways to restore earth's ecosystems if undertaken within a systems approach aimed at sustainable economic and environmental development.

Environmental justice research demonstrates that the negative environmental impacts of technology are commonly concentrated on 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. Ecological impact can thus exacerbate the economic and sociological effects of wealth disparities on human well-being by concentrating environmental injustice onto those who are less well off. Moreover, well-being research findings indicate that unfair economic and social inequality has a dampening effect on everyone's well-being, regardless of economic or social class.

In these respects, A/IS are no exception; they can be used in ways that either help or harm the ecological integrity of the planet. It may be fair to say that ecological health and human well-being will, increasingly, depend upon A/IS creators. It is imperative that A/IS creators and stakeholders find ways to use A/IS to do no harm and to reduce the environmental degradation associated with economic growth–while simultaneously identifying applications to restore the ecological health of the planet and thereby safeguarding the well-being of humans. For A/IS to reduce environmental degradation and promote well-being, it is required that not only A/IS creators act along such lines, but also that a systems approach is taken by all A/IS stakeholders to find solutions that safeguard human well-being with the understanding that human well-being is inextricable from healthy social, economic, and environmental systems.

Recommendations

A/IS creators need to recognize and prioritize the stewardship of the Earth's natural systems to promote human and ecological well-being. Specifically:

- Human well-being should be defined to encompass ecological health, access to nature, safe climate and natural environments, biosystem diversity, and other aspects of a healthy, sustainable natural environment.
- A/IS systems should be designed to use, support, and strengthen existing ecological sustainability standards with a certification or similar system, e.g., LEED, Energy Star, or Forest Stewardship Council. This directs automation and machine intelligence to follow the principle of doing no harm and to safeguard environmental, social, and economic systems.
- A/IS creators should prioritize doing no harm to the Earth's natural systems, both intended and unintended harm.
- A committee should be convened to issue findings on ways in which A/IS can be used by business, NGOs, and governmental agencies to promote stewardship and restoration of natural systems while reducing the harmful impact of economic development on ecological sustainability and environmental justice.

Well-being

Further Resources

- D. Austin and M. Macauley. "Cutting Through Environmental Issues: Technology as a double-edged sword." The Brookings Institution, Dec. 2001 [Online]. Available: https://www.brookings.edu/articles/cutting-through-environmental-issues-technology-as-a-double-edged-sword/. [Accessed Dec. 1, 2018].
- J. Newton, *Well-being and the Natural Environment: An Overview of the Evidence.* August 20, 2007.
- P. Dasgupta, Human Well-Being and the Natural Environment. Oxford, U.K.: Oxford University Press, 2001.
- R. Haines-Young and M. Potschin. "The Links Between Biodiversity, Ecosystem Services and Human Well-Being," in *Ecosystem Ecology: A New Synthesis*, D. Raffaelli, and C. Frid, Eds. Cambridge, U.K.: Cambridge University Press, 2010.
- S. Hart, *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' of the Holy Father Francis On the Care for Our Common Home. May 24, 2015.
- "Environment," The 14th Dalai Lama. Accessed Dec. 9, 2018. https://www.dalailama.com/messages/environment.
- Why Islam.org, Environment and Islam, 2018.

Issue: Human rights law is related to, but distinct from, the pursuit of well-being. Incorporating a human-rights framework as an essential basis for A/IS creators means A/IS creators honor existing law as part of their well-being analysis and implementation.

Background

International human rights law has been firmly established for decades in order to protect various guarantees and freedoms as enshrined in charters such as the United Nations' Universal Declaration of Human Rights and the Council of Europe's Convention on Human Rights. In 2018, the Toronto Declaration on machine learning standards was released, calling on both governments and technology companies to ensure that algorithms respect basic principles of equality and non-discrimination. The Toronto Declaration sets forth an obligation to prevent machine learning systems from discriminating, and in some cases violating, existing human rights law.

Well-being initiatives are typically undertaken for the sake of public interest. However, any metric, including well-being metrics, can be misused to justify human rights violations. Encampment and mistreatment of refugees and ethnic cleansing undertaken to preserve a nation's culture (an aspect of well-being) is one example. Imprisonment or assassination of journalists or researchers to ensure the stability

Well-being

of a government is another. The use of well-being metrics to justify human rights violations is an unconscionable perversion of the nature of any well-being metric. It should be noted that these same practices happen today in relation to GDP. For instance, in 2012, according to the International Labour Organization (ILO), approximately 21 million people are victims of forced labor (slavery), representing 9% to 56% of GDP income for various countries. These clear human rights violations, from sex trafficking and use of children in armies, to indentured farming or manufacturing labor, can increase a country's GDP while obviously harming human well-being.

Well-being metrics are designed to measure the efficacy of efforts related to individual and societal flourishing. Well-being as a value complements justice, equality, and freedom. Well-designed application of well-being considerations by A/IS creators should not displace other issues of human rights or ethical methodologies, but rather complement them.

Recommendation

A human rights framework should represent the floor, and not the ceiling, for the standards to which A/IS creators must adhere. Developers and users of well-being metrics should be aware these metrics will not always adequately address human rights.

Further Resources

- United Nations Universal Declaration of Human Rights, 1948.
- Council of Europe's Convention on Human Rights, 2018.
- International Labor Organization (ILO) Declaration on Fundamental Principles and Rights at Work, 1998.
- The regularly updated University of Minnesota Human Rights Library provides a wealth of material on human rights laws, its history, and the organizations engaged in promoting them.
- The Oxford Human Rights Hub reports on how and why technologies surrounding artificial intelligence raise human rights issues.

Section 2—Implementing Well-being Metrics for A/IS Creators

A key challenge for A/IS creators in realizing the benefits of well-being metrics is how to best incorporate them into their work. This section explores current best thinking on how to make this happen.

Issue: How can A/IS creators incorporate well-being into their work?

Background

Without practical ways of incorporating well-being metrics to guide, measure, and monitor impact, A/IS will likely lack fall short of its potential to avoid harm and promote well-being. Incorporating well-being thinking into typical organizational processes of design, prototyping, marketing, etc., suggests a variety of adaptations.

Organizations and A/IS creators should consider clearly defining the type of A/IS product or service that they are developing, including articulating its intended stakeholders and uses. By defining typical uses, possible uses, and finally unacceptable uses of the technology, creators will help to spell out the context of well-being. This can help to identify possible harms and risks given the different possible uses and end users, as well as intended and unintended positive consequences.

Additionally, internal and external stakeholders should be extensively consulted to ensure that impacts are thoroughly considered through an iterative and learning stakeholder engagement process. After consultation, A/IS creators should select appropriate well-being indicators based on the possible scope and impact of their A/IS product or service. These well-being indicators can be drawn from mainstream sources and models and adapted as necessary. They can be used to engage in pre-assessment of the intended user population, projection of possible impacts, and post-assessment. Development of a well-being indicator measurement plan and relevant data infrastructure will support a robust integration of well-being. A/IS models can also be trained to explicitly include well-being indicators as subgoals.

Data and discussions on well-being impacts can be used to suggest improvements and modifications to existing A/IS products and services throughout their lifecycle. 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, increase their sense of isolation, and lead to an overall decrease in their well-being. Therefore, even though a product modification may increase well-being according to one indicator or set of

Well-being

A/IS stakeholders, it does not mean that this modification should automatically be adopted.

Finally, organizational processes can be modified to incorporate the above strategies. Appointment of an organizational lead person for well-being impacts, e.g., a well-being lead, ombudsman, or officer can help to facilitate this effort.

Recommendation

A/IS creators should adjust their existing development, marketing, and assessment cycles to incorporate well-being concerns throughout their processes. This includes identification of an A/IS lead ombudsperson or officer; identification of stakeholders and end users; determination of possible uses, harm and risk assessment; robust stakeholder engagement; selection of well-being indicators; development of a well-being indicator measurement plan; and ongoing improvement of A/IS products and services throughout the lifecycle.

Further Resources

- Peter Senge and the Learning Organization - (synopsis) Purdue University
- Stakeholder Engagement: A Good Practice Handbook for Companies Doing Business in Emerging Markets. International Finance Corporation, May 2007.
- Global Reporting Initiative
- GNH Certification, Centre for Bhutan and GNH Studies, 2018.
- J. Helliwell, R. Layard, and J. Sachs, Eds., "The Objective Benefits of Subjective Well-Being," in World Happiness Report 2013. New York: UN Sustainable Development Solutions Network, pp. 54-79, 2013.
- Global Happiness and Well-being Policy Report by the Global Happiness Council, 2018.

Issue: How can A/IS creators influence A/IS goals to ensure well-being, and what can A/IS creators learn or borrow from existing models in the well-being and other arenas?

Background

Another way to incorporate considerations of well-being is to include well-being measures in the development, goal setting, and training of the A/IS systems themselves.

Identified metrics of well-being could be formulated as auxiliary objectives of the A/IS. As these auxiliary well-being objectives will be only a subset of the intended goals of the system, the architecture will need to balance multiple objectives. Each of these auxiliary objectives may be expressed as a goal, set of rules, set of values, or as a set of preferences, which can be weighted and combined using established methodologies from intelligent systems engineering.

Well-being

For example, an educational A/IS tool could not only optimize learning outcomes, but also incorporate measures of student social and emotional education, learning, and thriving.

A/IS-related data relates both to the individual—through personalized algorithms, in conjunction with affective sensors measuring and influencing emotion, and other aspects of individual well-being—and to society as large data sets representing aggregate individual subjective and objective data. As the exchange of this data becomes more widely available via establishing tracking methodologies, the data can be aligned within A/IS products and services to increase human well-being. For example, 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 Pepper units (and ostensibly their owners) benefit as well. As long as this data exchange happens with the predetermined consent of the robots' owners, this innovation in real time model can be emulated for the large-scale aggregation of information relating to existing well-being metrics.

A/IS creators can also help to operationalize well-being metrics by providing stakeholders with reports on the expected or actual outcomes of the A/IS and the values and objectives embedded in the systems. This transparency will help creators, users, and third parties assess the state of well-being produced by A/IS and make improvements in A/IS. In addition, A/IS creators should consider allowing end users to layer on their own preferences, such as allowing users to limit their use of an A/IS product if it leads to increased sustained stress levels, sustained isolation, development of unhealthy habits, or other decreases to well-being.

Incorporating well-being goals and metrics into broader organizational values and processes would support the use of well-being metrics as there would be institutional support. A key factor in industrial, corporate, and societal progress is cross-dissemination of concepts and models from one industry or field to another. To date, a number of successful concepts and models exist in the fields of sustainability, economics, industrial design and manufacturing, architecture and urban development, and governmental policy. These concepts and models can provide a foundation for building a metrics standard and the use of well-being metrics by A/IS creators, from conception and design to marketing, product updates, and improvements to the user experience.

Recommendation

Create technical standards for representing goals, metrics, and evaluation guidelines for well-being metrics and their precursors and components within A/IS that include:

- Ontologies for representing technological requirements.
- A testing framework for validating adherence to well-being metrics and ethical principles such as IEEE P7010™ Standards Project for Well-being Metric for Autonomous and Intelligent Systems.

Well-being

- The exploration of models and concepts listed above as well as others as a basis for a well-being metrics standard for A/IS creators. *(See page 191, Additional Resources: Additional Resources: Standards Development Models and Frameworks)*
- The development of a well-being metrics standard for A/IS that encompasses an understanding of well-being as holistic and interlinked to social, economic, and ecological systems.

Further Resources

- A.F.T 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
- R. A. Calvo, and D. Peters. Positive Computing: Technology for Well-Being and Human Potential. Cambridge MA: MIT Press, 2014.
- Y. Collette, and P. Slarry. Multiobjective Optimization: Principles and Case Studies (Decision Engineering Series). Berlin, Germany: Springer, 2004. doi: 10.1007/978-3-662-08883-8.
- J. Greene, 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.
- L. Li, I. Yevseyeva, V. Basto-Fernandes, H. Trautmann, N. Jing, and M. Emmerich,"Building and Using an Ontology of Preference-Based Multiobjective Evolutionary Algorithms." In 9th International Conference on Evolutionary Multi-Criterion Optimization—Volume 10173 (EMO 2017), H. Trautmann, G. Rudolph, K. Klamroth, O. Schütze, M. Wiecek, Y. Jin, and C. Grimme, Eds., Vol. 10173. Springer-Verlag, Berlin, Heidelberg, 406-421, 2017.
- PositiveSocialImpact: Empowering people, organizations and planet with information and knowledge to make a positive impact to sustainable development, 2017.
- D.K. Ura, Bhutan's Gross National Happiness Policy Screening Tool.

Issue: Decision processes for determining relevant well-being indicators through stakeholder deliberations need to be established.

Background

A/IS stakeholder involvement is necessary to determine relevant well-being indicators, for a number of reasons:

- "Well-being" will be defined differently by different groups affected by A/IS. The most relevant indicators of well-being may vary according to country, with concerns of wealthy nations being different than those of low- and middle-income countries. Indicators may vary based on geographical region or unique circumstances. The indicators may also be different across social groups, including gender, race, ethnicity, and disability status.
- Common indicators of well-being include satisfaction with life, healthy life expectancy,

Well-being

economic standard of living, trust in government, social support, perceived freedom to make life decisions, income equality, access to education, and poverty rates. Applying them in particular settings necessarily requires judgment, to ensure that assessments of well-being are in fact meaningful in context and reflective of the life circumstances of the diverse groups in question.

- Not all aspects of well-being are easily quantifiable. The importance of hard-to-quantify aspects of well-being is most likely to become apparent through interaction with those more directly affected by A/IS in specific settings.
- Engineers and corporate employees frequently misunderstand stakeholders' needs and expectations, especially when the stakeholders are very different from them in terms of educational and cultural background, social location, and/or economic status.

The processes through which stakeholders become involved in determining relevant well-being indicators will affect the quality of the indicators selected and assessed. Stakeholders should be empowered to define well-being, assess the appropriateness of existing indicators and propose new ones, and highlight context-specific factors that bear on issues of well-being, whether or not the issues have been recognized previously or are amenable to measurement. Interactive, open-ended discussions or deliberations among a wide variety of stakeholders and system designers are more likely to yield robust, widely-shared understandings of well-being and how to measure it in context. Closed-ended or over-determined methods for soliciting stakeholder input are likely to miss relevant information that system designers have not anticipated.

A process of stakeholder engagement and deliberation is one model for collective decision-making. Parties in such deliberation come together as equals. Their goal is to set aside their immediate, personal interests in order to think together about the common good. Participants in a stakeholder engagement and deliberation learn from one another's perspectives and experiences.

In the real world, stakeholder engagement and deliberation may run into the following challenges:

- Individuals with more education, power, or higher social status may—intentionally or unintentionally—dominate the discussion, undermining their ability to learn from less powerful participants.
- Topics may be preemptively ruled "out of bounds", to the detriment of collective problem-solving. An example would be if, in a deliberation on well-being and A/IS, participants were told that worries about the costs of health insurance were unrelated to A/IS and thus could not be discussed.
- Engineers and scientists may claim authority over technical issues and be willing to deliberate only on social issues, obscuring the ways that technical and social issues are intertwined.
- Less powerful groups may be unable to keep more powerful ones "at the table" when discussions get contentious, and vice versa.
- Participants may not agree on who can legitimately be involved in the conversation. For example, the consensual spirit of deliberation is often used as a justification for excluding activists and others who already hold a position on the issue.

Well-being

Stakeholder engagement and deliberative processes can be effective when:

- Their design is guided by experts or practitioners who are experienced in deliberation models.
- Deliberations are facilitated by individuals sensitive to issues of power and are skilled in mediating deliberation sessions.
- Less powerful actors participate with the help of allies who can amplify their voices.
- More powerful actors participate with an awareness of their own power and make a commitment to listen with humility, curiosity, and open-mindedness.
- Deliberations are convened by institutions or individuals who are trusted and respected by all parties and who hold all actors accountable for participating constructively.

Ethically aligned design of A/IS would be furthered by thoughtfully constructed, context-specific deliberations on well-being and the best indicators for assessing it.

Recommendation

Appoint a lead team or person, "leads", to facilitate stakeholder engagement and to serve as a resource for A/IS creators who use stakeholder-based processes to establish well-being indicators. Specifically:

- Leads should solicit and collect lessons learned from specific applications of stakeholder engagement and deliberation in order to continually refine its guidance.
- When determining well-being indicators, the leads should enlist the help of experts in public participation and deliberation. With expert guidance, facilitators can provide guidance for how to: take steps to mitigate the effects of unequal power in deliberative processes; incorporate appropriately trained facilitators and coaching participants in deliberations; recognize and curb disproportionate influence by more-powerful groups; use techniques to maximize the voices of less-powerful groups.
- Leads should use their convening power to bring together A/IS creators and stakeholders, including critics of A/IS, for deliberations on well-being indicators, impacts, and other considerations for specific contexts and settings. Leads' involvement would help bring actors to the table with a balance of power and encourage all actors to remain in conversation until robust, mutually agreeable definitions are found.

Further Resources

- D. E. Booher and J. E. Innes. Planning with Complexity: An Introduction to Collaborative Rationality for Public Policy. London: Routledge, 2010.
- J. A. Leydens and J. C. Lucena. Engineering Justice: Transforming Engineering Education and Practice. Wiley-IEEE Press, 2018.
- G. Ottinger. Assessing Community Advisory Panels: A Case Study from Louisiana's Industrial Corridor. Center for Contemporary History and Policy, 2008.
- Expert and Citizen Assessment of Science and Technology (ECAST) Network

Well-being

Issue: There are insufficient mechanisms to foresee and measure negative impacts, and to promote and safeguard positive impacts of A/IS.

Background

A/IS technologies present great opportunity for positive change in every aspect of society. However, they can—by design or unintentionally—cause harm as well. While it is important to consider and make sense of possible benefits, harms, and trade-offs, it is extremely challenging to foresee all of the relevant, direct, and secondary impacts.

However, it is prudent to review case studies of similar products and the impacts they have had on well-being, as well as to consider possible types of impacts that could apply. Issues to consider include:

- Economic and labor impacts, including labor displacement, unemployment, and inequality,
- Accountability, transparency, and explainability,
- Surveillance, privacy, and civil liberties,
- Fairness, ethics, and human rights,
- Political manipulation, deception, "nudging", and propaganda,
- Human physical and psychological health,
- Environmental impacts,
- Human dignity, autonomy, and human vs. A/IS roles,
- Security, cybersecurity, and autonomous weapons, and
- Existential risk and super intelligence.

While this is a partial list, it is important to be aware of and reflect on possible and actual cases. For example:

- A prominent concern related to A/IS is of labor displacement and economic and social impacts at an individual and a systems level. A/IS technologies designed to replicate human tasks, behavior, or emotion have the potential to increase or decrease human well-being. These systems could complement human work and increase productivity, wages, and leisure time; or they could be used to supplement and displace human workers, leading to unemployment, inequality, and social strife. It is important for A/IS creators to think about possible uses of their technology and whether they want to encourage or design in restrictions in light of these impacts.
- Another example relates to manipulation. Sophisticated manipulative technologies utilizing A/IS can restrict the fundamental freedom of human choice by manipulating humans who consume content without them recognizing the extent of the manipulation. Software platforms are moving from targeting and customizing content to much more powerful and potentially harmful "persuasive computing" that leverages psychological data and methods. While these approaches may be effective in encouraging use of a product, they may come at significant psychological and social costs.
- A/IS may 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 intelligence from artificial intelligence, there is a significant risk

Well-being

that unscrupulous operators will abuse the technology for unethical commercial or outright criminal purposes. 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 and via multiple mediums.

A potential entry point for exploring these unintended consequences is computational sustainability.

Computational-Sustainability.org defines the term as an "interdisciplinary field that aims to apply techniques from computer science, information science, operations research, applied mathematics, and statistics for balancing environmental, economic, and societal needs for sustainable development". The Institute of Computational Sustainability states that the intent of computational sustainability is provide "computational models for a sustainable environment, economy, and society". Examples of applied computational sustainability can be seen in the Stanford University Engineering Department's course in computational sustainability presentation. Computational sustainability technologies designed to increase social good could also be tied to existing well-being metrics.

Recommendation

- To avoid potential negative, unintended consequences, and secure and safeguard positive impacts, A/IS creators, end-users, and stakeholders should be aware of possible well-being impacts when designing, using, and monitoring A/IS systems. This includes being aware of existing cases and possible areas of impact, measuring impacts on well-being outcomes, and developing regulations to promote beneficent uses of A/IS. Specifically:
- A/IS creators should protect human dignity, autonomy, rights, and well-being of those directly and indirectly affected by the technology. As part of this effort, it is important to include multiple stakeholders, minorities, marginalized groups, and those often without power or a voice in consultation.
- Policymakers, regulators, monitors, and researchers should consider issuing guidance on areas such as A/IS labor and the proper role of humans vs. A/IS in work transparency, trust, and explainability; manipulation and deception; and other areas that emerge.
- Ongoing literature review and analysis should be performed by research and other communities to curate and aggregate information on positive and negative A/IS impacts, along with demonstrated approaches to realize positive ones and ameliorate negative ones.
- A/IS creators working toward computational sustainability should integrate well-being concepts, scientific findings, and indicators into current computational sustainability models. They should work with well-being experts, researchers, and practitioners to conduct research and develop and apply models in A/IS development that prioritize and increase human well-being.

Well-being

- Cross-pollination should be developed between computational sustainability and well-being professionals to ensure integration of well-being into computational sustainability frameworks, and vice versa. Where feasible and reasonable, do the same for conceptual models such as doughnut economics and systems thinking.

Further Resources

- AI Safety Research by The Future of Life Institute
- D. Helbing, et al. "Will Democracy Survive Big Data and Artificial Intelligence?" *Scientific American*, February 25, 2017.
- J. L. Schenker, "Can We Balance Human Ethics with Artificial Intelligence?" *Techonomy,* January 23, 2017.
- M. Bulman, "EU to Vote on Declaring Robots To Be 'Electronic Persons'." *Independent,* January 14, 2017.
- N. Nevejan, for the European Parliament. "European Civil Law Rules in Robotics." October 2016.
- University of Oxford. "Social media manipulation rising globally, new report warns," https://phys.org/news/2018-07-social-media-globally.html. July 20, 2018.
- "The AI That Pretends To Be Human," *LessWrong* blog post, February 2, 2016.
- C. Chan, "Monkeys Grieve When Their Robot Friend Dies." *Gizmodo*, January 11, 2017.
- Partnership on AI, "AI, Labor, and the Economy" Working Group launches in New York City," https://www.partnershiponai.org/aile-wg-launch/. April 25, 2018.
- C.Y. Johnson, "Children can be swayed by robot peer pressure,study says," The Washington Post, August 15, 2018. [Online]. Available: www.WashingtonPost.com. [Accessed 2018].

Further Resources for Computational Sustainability

- Stanford Engineering Department, Topics in Computational Sustainability Course Presentation, 2016.
- Computational Sustainability, Computational Sustainability: Computational Methods for a Sustainable Environment, Economy, and Society Project Summary.
- C. P. Gomes, "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.
- S.J. Gershman, E. J. Horvitz, and J. B. Tenenbaum. "Computational rationality: A converging paradigm for intelligence in brains, minds, and machines," *Science* vol. 349, no. 6245, pp. 273–278, July 2015.
- ACM Fairness, Accountability and Transparency Conference

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Well-being Committee

- **John C. Havens** (Co-Chair) – Executive Director, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems; Executive Director, The Council on Extended Intelligence; Author, *Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*
- **Laura Musikanski** (Co-Chair) – Executive Director at The Happiness Alliance—home of The Happiness Initiative & Gross National Happiness Index
- **Liz Alexander** – PhD Futurist
- **Anna Alexandrova** – Senior Lecturer in Philosophy of Science at Cambridge University and Fellow of Kings College
- **Christina Berkley** – Executive Coach to leaders in exponential technologies, cutting-edge science, and aerospace
- **Catalina Butnaru** – UK AI Ambassador for global community City.AI, and Founder of HAI, the first methodology for applications of AI in cognitive businesses
- **Celina Beatriz** – Project Director at the Institute for Technology & Society of Rio de Janeiro (ITS Rio)
- **Peet van Biljon** – Founder and CEO at BMNP Strategies LLC, advisor on strategy, innovation, and business transformation; Adjunct faculty at Georgetown University; Business ethics author
- **Amy Blankson** – Author of The Future of Happiness and Founder of TechWell, a research and consulting firm that aims to help organizations to create more positive digital cultures
- **Marc Böhlen** – Professor, University at Buffalo, Emerging Practices in Computational Media. www.realtechsupport.org
- **Rafael A. Calvo** – Professor and ARC Future Fellow at The University of Sydney. Co-author of Positive Computing: Technology for Well-Being and Human Potential
- **Rumman Chowdhury** – PhD Senior Principal, Artificial Intelligence, and Strategic Growth Initiative Responsible AI Lead, Accenture
- **Dr. Aymee Coget** – CEO and Founder of Happiness For HumanKind
- **Danny W. Devriendt** – Managing director of Mediabrands Dynamic (IPG) in Brussels, and the CEO of the Eye of Horus, a global think-tank for communication-technology related topics
- **Eimear Farrell** – Eimear Farrell, independent expert/consultant on technology and human rights (formerly at OHCHR)
- **Danit Gal** – Project Assistant Professor, Keio University; Chair, IEEE Standard P7009 on the Fail-Safe Design of Autonomous and Semi-Autonomous Systems

Well-being

- **Marek Havrda** – PhD Strategy Advisor, GoodAI
- **Andra Keay** – Managing Director of Silicon Valley Robotics, cofounder of Robohub
- **Dr. Peggy Kern** – Senior Lecturer, Centre for Positive Psychology at the University of Melbourne's Graduate School of Education
- **Michael Lennon** – Senior Fellow, Center for Excellence in Public Leadership, George Washington University; Co-Founder, Govpreneur.org; Principal, CAIPP.org (Consortium for Action Intelligence and Positive Performance); Member, Well-being Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems Committee
- **Alan Mackworth** – Professor of Computer Science, University of British Columbia; Former President, AAAI; Co-author of "Artificial Intelligence: Foundations of Computational Agents"
- **Richard Mallah** – Director of AI Project, Future of Life Institute
- **Fabrice Murtin** – Senior Economist, OECD Statistics and Data Directorate
- **Gwen Ottinger** – Associate Professor, Center for Science, Technology, and Society and Department of Politics, Drexel University; Director, Fair Tech Collective
- **Eleonore Pauwels** – Research Fellow on AI and Emerging Cybertechnologies, United Nations University (NY) and Director of the AI Lab, Woodrow Wilson International Center for Scholars (DC)
- **Venerable Tenzin Priyadarshi** – MIT Media Lab, Director, Ethics Initiative
- **Gideon Rosenblatt** – Writer, focused on work and the human experience in an era of machine intelligence, at The Vital Edge
- **Daniel Schiff** – PhD Student, Georgia Institute of Technology; Chair, Sub-Group for Autonomous and Intelligent Systems Implementation, IEEE P7010™ Standards Project for Well-being Metric for Autonomous and Intelligent Systems
- **Madalena Sula** – Undergraduate student of Electrical and Computer Engineering, University of Thessaly, Greece, x-PR Manager of IEEE Student Branch of University of Thessaly, Data Scientist & Business Analyst in a multinational company
- **Vincent Siegerink** – Analyst, OECD Statistics and Data Directorate
- **Andy Townsend** – Emerging and Disruptive Technology, PwC UK
- **Andre Uhl** – Research Associate, Director's Office, MIT Media Lab
- **Ramón Villasante** – Founder of PositiveSocialImpact. Software designer, engineer, CTO & CPO in EdTech for sustainable development, social impact and innovation
- **Sarah Villeneuve** – Policy Analyst; Member, IEEE P7010™ Standards Project for Well-being Metric for Autonomous and Intelligent Systems.

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

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. Emotions are one mechanism that humans evolved to accomplish what needs to be done in the time available with the information at hand—to satisfice. Emotions are not an impediment to rationality; arguably they are integral to rationality in humans. Humans create and respond to both positive and negative emotional influence as they coordinate their actions with other individuals to create societies. Autonomous and intelligent systems (A/IS) are being designed to simulate emotions in their interactions with humans in ways that will alter our societies.

A/IS should be used to help humanity to the greatest extent possible in as many contexts as are appropriate. While A/IS have tremendous potential to effect positive change, there is also potential that artifacts used in society could cause harm either by amplifying, altering, or even dampening human emotional experience. Even rudimentary versions of synthetic emotions, such as those already in use within nudging systems, have already altered the perception of A/IS by the general public and public policy makers.

This chapter of *Ethically Aligned Design* addresses issues related to emotions and emotion-like control in interactions between humans and design of A/IS. We have put forward recommendations on a variety of topics: considering how affect varies across human cultures; the particular problems of artifacts designed for caring and private relationships; considerations of how intelligent artifacts may be used for "nudging"; how systems can support human flourishing; and appropriate policy interventions for artifacts designed with inbuilt affective systems.

Document Sections

Section 1—Systems Across Cultures

Issue: Should affective systems interact using the norms for verbal and nonverbal communication consistent with the norms of the society in which they are embedded?

Background

Individuals around the world express intentions differently, including the ways that they make eye contact, use gestures, or interpret silence. These particularities are part of an individual's and a society's culture and are incorporated into their affective systems in order to convey the intended message. To ensure that the emotional systems of autonomous and intelligent systems foster effective communication within a specific culture, an understanding of the norms/values of the community where the affective system will be deployed is essential.

Recommendations

1. A well-designed affective system will have a set of essential norms, specific to its intended cultural context of use, in its knowledge base. Research has shown that A/IS technologies can use at least five types of cues to simulate social interactions.
2. These include: physical cues such as simulated facial expressions, psychological cues such as simulated humor or other emotions, use of language, use of social dynamics like taking turns, and through social roles such as acting as a tutor or medical advisor. Further examples are listed below:

    a. Well-designed affective systems will use language with affective content carefully and within the contemporaneous expectations of the culture. An example is small talk. Although small talk is useful for establishing a friendly rapport in many communities, some communities see people that use small talk as insincere and hypocritical. Other cultures may consider people that do not use small talk as unfriendly, uncooperative, rude, arrogant, or ignorant. Additionally, speaking with proper vocabulary, grammar, and sentence structure may contrast with the typical informal interactions between individuals. For example, the latest trend, TV show, or other media may significantly influence what is viewed as appropriate vocabulary and interaction style.

    b. Well-designed affective systems will recognize that the amount of personal space (proxemics) given by individuals in an important part of culturally specific

Affective Computing

human interaction. People from varying cultures maintain, often unknowingly, different spatial distances between themselves to establish smooth communication. Crossing these limits may require explicit or implicit consent, which A/IS must learn to negotiate to avoid transmitting unintended messages.

c. Eye contact is an essential component for culturally sensitive social interaction. For some interactions, direct eye contact is needed but for others it is not essential and may even generate misunderstandings. It is important that A/IS be equipped to recognize the role of eye contact in the development of emotional interaction.

d. Hand gestures and other non-verbal communication are very important for social interaction. Communicative gestures are culturally specific and thus should be used with caution in cross-cultural situations. The specificity of physical communication techniques must be acknowledged in the design of functional 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.

e. Humans use facial expressions to detect emotions and facilitate communication. Facial expressions may not be universal across cultures, however, and A/IS trained with a dataset from one culture may not be readily usable in another culture. Well-developed A/IS will be able to recognize, analyze, and even display facial expressions essential for culturally specific social interaction.

3. Engineers should consider the need for cross-cultural use of affective systems. Well-designed systems will have options innate to facilitate flexibility in cultural programming. Mechanisms to enable and disable culturally specific "add-ons" should be considered an essential part of A/IS development.

Further Resources

- G. Cotton, "Gestures to Avoid in Cross-Cultural Business: In Other Words, 'Keep Your Fingers to Yourself!'" *Huffington Post,* June 13, 2013.
- "Paralanguage Across Cultures," Sydney, Australia: Culture Plus Consulting, 2016.
- G. Cotton, *Say Anything to Anyone, Say Anything to Anyone, Anywhere: 5 Keys to Successful Cross-Cultural Communication*. Hoboken, NJ: Wiley, 2013.
- D. Elmer, *Cross-Cultural Connections: Stepping Out and Fitting In Around the World*. Westmont, IL: InterVarsity Press, 2002.
- B. J. Fogg, Persuasive Technology. *Ubiquity*, December 2, 2002.
- A. McStay, Emotional AI: The Rise of Empathic Media. London: Sage, 2018.
- M. Price, "Facial Expressions—Including Fear—May Not Be as Universal as We Thought." *Science,* October 17, 2016.

Affective Computing

Issue: It is presently unknown whether long-term interaction with affective artifacts that lack cultural sensitivity could alter human social interaction.

Background

Systems that do not have cultural knowledge incorporated into their knowledge base may or may not interact effectively with humans for whom emotion and culture are significant. Given that interaction with A/IS may affect individuals and societies, it is imperative that we carefully evaluate mechanisms to promote beneficial affective interaction between humans and A/IS. Humans often use mirroring in order to understand and develop their norms for behavior. Certain machine learning approaches also address improving A/IS interaction with humans through mirroring human behavior. Thus, we must remember that learning via mirroring can go in both directions and that interacting with machines has the potential to impact individuals' norms, as well as societal and cultural norms. If affective artifacts with enhanced, different, or absent cultural sensitivity interact with impressionable humans this could alter their responses to social and cultural cues and values. The potential for A/IS to exert cultural influence in powerful ways, at scale, is an area of substantial concern.

Recommendations

1. Collaborative research teams must research the effects of long-term interaction of people with affective systems. This should be done using multiple protocols, disciplinary approaches, and metrics to measure the modifications of habits, norms, and principles as well as careful evaluation of the downstream cultural and societal impacts.

2. Parties responsible for deploying affective systems into the lives of individuals or communities should be trained to detect the influence of A/IS, and to utilize mitigation techniques if A/IS effects appear to be harmful. It should always be possible to shut down harmful A/IS.

Further Resources

- T. Nishida and C. Faucher, Eds., Modelling Machine Emotions for Realizing Intelligence: Foundations and Applications. Berlin, Germany: Springer-Verlag, 2010.

- D. J. Pauleen, 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*, vol. 17,)pp. 1–36, 2006. J. Bielby, "Comparative Philosophies in Intercultural Information Ethics." *Confluence: Online Journal of World Philosophies* 2, no. 1, pp. 233–253, 2015.

- J. Bryson, "Why Robot Nannies Probably Won't Do Much Psychological Damage." A commentary on an article by N. Sharkey

Affective Computing

and A. Sharkey, *The Crying Shame of Robot Nannies.* Interaction Studies, vol. 11, no. 2 pp. 161–190, July 2010.

- A. Sharkey, and N. Sharkey, "Children, the Elderly, and Interactive Robots." *IEEE Robotics & Automation Magazine,* vol.18, no. 1, pp. 32–38, March 2011.

Issue: When affective systems are deployed across cultures, they could adversely affect the cultural, social, or religious values of the community in which they interact.

Background

Some philosophers argue that there are no universal ethical principles and 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 they are embedded. How systems should effectively reflect the values of the designers or the users of affective systems is not a settled discussion. There is general agreement that developers of affective systems should acknowledge that the systems should reflect the values of those with whom 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, and that end-user values should be actively considered. Differences between affective systems and societal values may generate conflict situations producing undesirable results, e.g., gestures or eye contact being misunderstood as rude or threatening. Thus, affective systems should adapt to reflect the values of the community and individuals where they will operate in order to avoid misunderstanding.

Recommendations

Assuming that well-designed affective systems have a minimum subset of configurable norms incorporated in their knowledge base:

1. Affective systems should have capabilities to identify differences between the values they are designed with and the differing values of those with whom the systems are interacting.

2. Where appropriate, affective systems will adapt accordingly over time to better fit the norms of their users. As societal values change, there needs to be a means to detect and accommodate such cultural change in affective systems.

3. Those actions undertaken by an affective system that are most likely to generate an emotional response should be designed to be easily changed in appropriate ways by the user without being easily hacked by actors with malicious intentions. Similar to how software today externalizes the language and vocabulary to be easily changeable based on location, affective systems should externalize some of the core aspects of their actions.

Further Resources

- J. Bielby, "Comparative Philosophies in Intercultural Information Ethics." *Confluence: Online Journal of World Philosophies* 2, no. 1, pp. 233–253, 2015.
- M. Velasquez, 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.
- T. Donaldson, "Values in Tension: Ethics Away from Home Away from Home." *Harvard Business Review.* September– October 1996.

Section 2—When Systems Care

Issue: Are moral and ethical boundaries crossed when the design of affective systems allows them to develop intimate relationships with their users?

Background

There are many robots in development or production designed to focus on intimate care of children, adults, and the elderly[2]. While robots capable of participating fully in intimate relationships are not currently available, the potential use of such robots routinely captures the attention of the media. It is important that professional communities, policy makers, and the general public participate in development of guidelines for appropriate use of A/IS in this area. Those guidelines should acknowledge fundamental human rights to highlight potential ethical benefits and risks that may emerge, if and when affective systems interact intimately with users.

Among the many areas of concern are the representation of care, embodiment of caring A/IS, and the sensitivity of data generated through intimate and caring relationships with A/IS. The literature suggests that there are some potential benefits to individuals and to society from the incorporation of caring A/IS, along with duly cautionary notes concerning the possibility that these systems could negatively impact human-to-human intimate relations[3].

Recommendations

As this technology develops, it is important to monitor research into the development of intimate relationships between A/IS and humans. Research should emphasize any technical and

Affective Computing

normative developments that reflect use of A/IS in positive and therapeutic ways while also creating appropriate safeguards to mitigate against uses that contribute to problematic individual or social relationships:

1. Intimate systems must not be designed or deployed in ways that contribute to stereotypes, gender or racial inequality, or the exacerbation of human misery.

2. Intimate systems must not be designed to explicitly engage in the psychological manipulation of the users of these systems unless the user is made aware they are being manipulated and consents to this behavior. Any manipulation should be governed through an opt-in system.

3. Caring A/IS should be designed to avoid contributing to user isolation from society.

4. Designers of affective robotics must publicly acknowledge, for example, within a notice associated with the product, that these systems can have side effects, such as interfering with the relationship dynamics between human partners, causing attachments between the user and the A/IS that are distinct from human partnership.

5. Commercially marketed A/IS for caring applications should not be presented to be a person in a legal sense, nor marketed as a person. Rather its artifactual, that is, 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 Section 4, Systems Supporting Human Potential.

6. Existing laws regarding personal imagery need to be reconsidered in light of caring A/IS. In addition to other ethical considerations, it will also be necessary to establish conformance with local laws and mores in the context of caring A/IS systems.

Further Resources

- M. Boden, J. Bryson, D. Caldwell, K. Dautenhahn, L. Edwards, S. Kember, P. Newman, V. Parry, G. Pegman, T. Rodden and T. Sorrell, Principles of robotics: regulating robots in the real world. Connection Science, vol. 29, no. 2, pp. 124-129, April 2017.

- J. J. Bryson, M. E. Diamantis, and T. D. Grant, "Of, For, and By the People: The Legal Lacuna of Synthetic Persons." *Artificial Intelligence & Law*, vol. 25, no. 3, pp. 273–291, Sept. 2017.

- M. Scheutz, "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots," in *Robot Ethics: The Ethical and Social Implications of Robotics,* P. Lin, K. Abney, and G. Bekey, Eds., pp. 205. Cambridge, MA: MIT Press, 2011.

Section 3– 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 others?

Background

Manipulation can be defined as an exercise of influence by one person or group, with the intention to attempt to control or modify the actions of another person or group. Thaler and Sunstein (2008) call the tactic of subtly modifying behavior a "nudge[4]". Nudging mainly operates through the affective elements of a human rational system. Making use of a nudge might be considered appropriate in situations like teaching children, treating drug dependency, and in some healthcare settings. While nudges can be deployed to encourage individuals to express behaviors that have community benefits, a nudge could have unanticipated consequences for people whose backgrounds were not well considered in the development of the nudging system[5]. Likewise, nudges may encourage behaviors with unanticipated long-term effects, whether positive or negative, for the individual and/or society. The effect of A/IS nudging a person, such as potentially eroding or encouraging individual liberty, or expressing behaviors that are for the benefit others, should be well characterized in the design of A/IS.

Recommendations

1. Systematic analyses are needed that examine the ethics and behavioral consequences of designing affective systems to nudge human beings prior to deployment.
2. The user should be empowered, through an explicit opt-in system and readily available, comprehensible information, to recognize different types of A/IS nudges, regardless of whether they seek to promote beneficial social manipulation or to enhance consumer acceptance of commercial goals. The user should be able to access and check facts behind the nudges and then make a conscious decision to accept or reject a nudge. Nudging systems must be transparent, with a clear chain of accountability that includes human agents: data logging is required so users can know how, why, and by whom they were nudged.
3. A/IS nudging must not become coercive and should always have an opt-in system policy with explicit consent.
4. Additional protections against unwanted nudging must be put in place for vulnerable populations, such as children, or when informed consent cannot be obtained. Protections against unwanted nudging should be encouraged when nudges alter long-term behavior or when consent alone may not be a sufficient safeguard against coercion or exploitation.

Affective Computing

5. Data gathered which could reveal an individual or groups' susceptibility to a nudge or their emotional reaction to a nudge should not be collected or distributed without opt-in consent, and should only be retained transparently, with access restrictions in compliance with the highest requirements of data privacy and law.

Further Resources

- R. Thaler, and C. R. Sunstein, *Nudge: Improving Decision about Health, Wealth and Happiness*, New Haven, CT: Yale University Press, 2008.
- L. Bovens, "The Ethics of Nudge," in *Preference change: Approaches from Philosophy, Economics and Psychology*, T. Grüne-Yanoff and S. O. Hansson, Eds., Berlin, Germany: Springer, 2008 pp. 207–219.
- S. D. Hunt and S. Vitell. "A General Theory of Marketing Ethics." Journal of Macromarketing, vol.6, no. 1, pp. 5-16, June 1986.
- A. McStay, Empathic Media and Advertising: Industry, Policy, Legal and Citizen Perspectives (the Case for Intimacy), Big Data & Society, pp. 1-11, December 2016.
- J. de Quintana Medina 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.
- M. D. White, *The Manipulation of Choice. Ethics and Libertarian Paternalism.* New York: Palgrave Macmillan, 2013
- C.R. Sunstein, The Ethics of Influence: Government in the Age of Behavioral Science. New York: Cambridge, 2016
- M. Scheutz, "The Affect Dilemma for Artificial Agents: Should We Develop Affective Artificial Agents? " *IEEE Transactions on Affective Computing,* vol. 3, no. 4,pp. 424–433, Sept. 2012.
- A. Grinbaum, R. Chatila, L. Devillers, J.-G. Ganascia, C. Tessier and M. Dauchet. "Ethics in Robotics Research: CERNA Recommendations," *IEEE Robotics and Automation Magazine,* vol. 24, no. 3,pp. 139–145, Sept. 2017.
- "Designing Moral Technologies: Theoretical, Practical, and Ethical Issues" Conference July 10–15, 2016, Monte Verità, Switzerland.

Affective Computing

Issue: Governmental entities may potentially use nudging strategies, for example to promote the performance of charitable acts. Does the practice of nudging for the benefit of society, including nudges by affective systems, raise ethical concerns?

Background

A few scholars have noted a potentially controversial practice of the future: allowing a robot or another affective system to nudge a user for the good of society[6]. 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.

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.

Recommendations

1. 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 in the context of governmental actions.
2. 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

- J. Borenstein and R. Arkin, "Robotic Nudges: Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being Just Human Being." *Science and Engineering Ethics,* vol. 22, no. 1,pp. 31–46, Feb. 2016.
- J. Borenstein and R. Arkin. "Nudging for Good: Robots and the Ethical Appropriateness of Nurturing Empathy and Charitable Behavior ." *AI and Society,* vol. 32, no. 4, pp. 499–507, Nov. 2016.

Issue: Will A/IS nudging systems that are not fully relevant to the sociotechnical context in which they are operating cause behaviors with adverse unintended consequences?

Background

A well-designed nudging or suggestion system will have sophisticated enough technical capabilities for recognizing the context in which it is applying nudging actions. Assessment of the context requires perception of the scope or impact of the actions to be taken, the consequences of incorrectly or incompletely

Affective Computing

applied nudges, and acknowledgement of the uncertainties that may stem from long term consequences of a nudge[7].

Recommendations

1. Consideration should be given to the development of a system of technical licensing ("permits") or other certification from governments or non-governmental organizations (NGOs) that can aid users to understand the nudges from A/IS in their lives.

2. User autonomy is a key and essential consideration that must be taken into account when addressing whether affective systems should be permitted to nudge human beings.

3. Design features of an affective system that nudges human beings should include the ability to accurately distinguish between users, including detecting characteristics such as whether the user is an adult or a child.

4. Affective systems with nudging strategies should incorporate a design system of evaluation, monitoring, and control for unintended consequences.

Further Resources

- J. Borenstein and R. Arkin, "Robotic Nudges: Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being Just Human Being." *Science and Engineering Ethics,* vol. 22, no. 1, pp. 31–46, 2016.
- R. C. Arkin, M. Fujita, T. Takagi, and R. Hasegawa, "An Ethological and Emotional Basis for Human- Robot Interaction." *Robotics and Autonomous Systems,* vol. 42, no. 3–4 pp.191–201, March 2003.
- S. Omohundro "Autonomous Technology and the Greater Human Good." *Journal of Experimental and Theoretical Artificial Intelligence, vol.* 26, no. 3, pp. 303–315, 2014.

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 indicate that it can be acceptable when deception increases overall happiness. Given the diversity of views on ethics and the appropriateness of deception, should affective systems be designed to deceive? Does the non-consensual nature of deception restrict the use of A/IS in contexts in which deception may be required?

Affective Computing

Recommendations

It is necessary to develop recommendations regarding the acceptability of deception performed by A/IS, specifically with respect to when and under which circumstances, if any, it is appropriate.

1. In general, deception may be 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 or for 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 should be certified by an external authority, such as a licensing body or regulatory agency.

Further Resources

- R. C. Arkin, "Robots That Need to Mislead: Biologically-inspired Machine Deception." *IEEE Intelligent Systems* 27, no. 6, pp. 60–75, 2012.

- J. Shim and R. C. Arkin, "Other-Oriented Robot Deception: How Can a Robot's Deceptive Feedback Help Humans in HRI?" *Eighth International Conference on Social Robotics (ICSR 2016)*, Kansas, MO., November 2016.

- J. Shim 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.

- J. Shim 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.

Section 4—Systems Supporting Human Potential

Issue: Will extensive use of A/IS in society make our organizations more brittle by reducing human autonomy within organizations, and by replacing creative, affective, empathetic components of management chains?

Background

If human workers are replaced by A/IS, the possibility of corporations, governments, employees, and customers discovering new equilibria outside the scope of what the organizations' past leadership originally foresaw may be unduly limited. A lack of empathy based on shared needs, abilities, and disadvantages between organizations and customers causes disequilibria between the individuals and corporations and governments that exist to serve them. Opportunities for useful innovation may therefore be lost through automation. Collaboration requires enough commonality of collaborating intelligences to create empathy—the capacity to model the other's goals based on one's own.

According to scientists within several fields, autonomy is a psychological need. Without it, humans fail to thrive, create, and innovate. Ethically aligned design should support, not hinder, human autonomy or its expression.

Recommendations

1. It is important that human workers' interaction with other workers not always be intermediated by affective systems (or other technology) which may filter out autonomy, innovation, and communication.
2. Human points of contact should remain available to customers and other organizations when using A/IS.
3. Affective systems should be designed to support human autonomy, sense of competence, and meaningful relationships as these are necessary to support a flourishing life.
4. Even where A/IS are less expensive, more predictable, and easier to control than human employees, a core network of human employees should be maintained at every level of decision-making in order to ensure preservation of human autonomy, communication, and innovation.
5. Management and organizational theorists should consider appropriate use of affective and autonomous systems to enhance their business models and the efficacy of their workforce within the limits of the preservation of human autonomy.

Affective Computing

Further Resources

- J. J. Bryson, "Artificial Intelligence and Pro-Social Behavior," in *Collective Agency and Cooperation in Natural and Artificial Systems,* C. Misselhorn, Ed., pp. 281–306, Springer, 2015.
- D. Peters, R.A. Calvo, and R.M. Ryan, "Designing for Motivation, Engagement and Wellbeing in Digital Experience," *Frontiers in Psychology*– Human Media Interaction, vol. 9, pp 797, 2018.

Issue: Does the increased access to personal information about other members of our society, facilitated by A/IS, alter the human affective experience? Does this access potentially lead to a change in human autonomy?

Background

Theoretical biology tells us that we should expect increased communication—which A/IS facilitate—to increase group-level investment[8]. Extensive use of A/IS could change the expression of individual autonomy and in its place increase group-based identities. Examples of this sort of social alteration may include:

1. Changes in the scope of monitoring and control of children's lives by parents.
2. Decreased willingness to express opinions for fear of surveillance or long-term consequences of past expressions being used in changed temporal contexts.
3. Utilization of customers or other end users to perform basic corporate business processes such as data entry as a barter for lower prices or access, resulting potentially in reduced tax revenues.
4. Changes to the expression of individual autonomy could alter the diversity, creativity, and cohesiveness of a society. It may also alter perceptions of privacy and security, and social and legal liability for autonomous expressions.

Recommendations

1. Organizations, including governments, must put a high value on individuals' privacy and autonomy, including restricting the amount and age of data held about individuals specifically.
2. Education in all forms should encourage individuation, the preservation of autonomy, and knowledge of the appropriate uses and limits to A/IS[9].

Further Resources

- J. J. Bryson, "Artificial Intelligence and Pro-Social Behavior," in *Collective Agency and Cooperation in Natural and Artificial Systems,* C. Misselhorn, Ed., pp. 281–306, Springer, 2015.
- M. Cooke, "A Space of One's Own: Autonomy, Privacy, Liberty," *Philosophy & Social Criticism,* Vol. 25, no. 1, pp. 22–53, 1999.
- D. Peters, R.A. Calvo, R.M. Ryan, "Designing for Motivation, Engagement and Wellbeing in Digital Experience" *Frontiers in Psychology – Human Media Interaction*, vol. 9. pp 797, 2018.

- J. Roughgarden, M. Oishi and E. Akçay, "Reproductive Social Behavior: Cooperative Games to Replace Sexual Selection." *Science* 311, no. 5763, pp. 965–969, 2006.

Issue: Will use of A/IS adversely affect human psychological and emotional well-being in ways not otherwise foreseen?

Background

A/IS may be given unprecedented access to human culture and human spaces—both physical and intellectual. A/IS may communicate via natural language, may move with humanlike form, and may express humanlike identity, but they are not, and should not be regarded as, human. Incorporation of A/IS into daily life may affect human well-being in ways not yet anticipated. Incorporation of A/IS may alter patterns of trust and capability assessment between humans, and between humans and A/IS.

Recommendations

1. Vigilance and robust, interdisciplinary, on-going research on identifying situations where A/IS affect human well-being, both positively and negatively, is necessary. Evidence of correlations between the increased use of A/IS and positive or negative individual or social outcomes must be explored.
2. Design restrictions should be placed on the systems themselves to avoid machine decisions that may alter a person's life in unknown ways. Explanations should be available on demand in systems that may affect human well-being.

Further Resources

- K. Kamewari, 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,* vol. 20, no. 2, pp. 303–320, 2005.
- R.A. Calvo and D. Peters, Positive Computing: Technology for Wellbeing and Human Potential. Cambridge, MA: MIT Press, 2014.

Section 5—Systems with Synthetic Emotions

Issue: Will deployment of synthetic emotions into affective systems increase the accessibility of A/IS? Will increased accessibility prompt unforeseen patterns of identification with A/IS?

Background

Deliberately constructed emotions are designed to create empathy between humans and artifacts, which may be useful or even essential for human-A/IS collaboration. Synthetic emotions are essential for humans to collaborate with the A/IS but can also lead to failure to recognize that synthetic emotions can be compartmentalized and even entirely removed. Potential consequences for humans include different patterns of bonding, guilt, and trust, whether between the human and A/IS or between other humans. There is no coherent sense in which A/IS can be made to suffer emotional loss, because any such affect, even if possible, could be avoided at the stage of engineering, or reengineered. As such, it is not possible to allocate moral agency or responsibility in the senses that have been developed for human emotional bonding and thus sociality.

Recommendations

1. Commercially marketed A/IS should not be persons in a legal sense, nor marketed as persons. Rather their 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 systems should be available to inspection by responsible parties but may be withdrawn for operational needs.

Further Resources

- R. C. Arkin, P. Ulam and A. R. Wagner, "Moral Decision-making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust and Deception," *Proceedings of the IEEE,* vol. 100, no. 3, pp. 571–589, 2012.
- R. Arkin, M. Fujita, T. Takagi and R. Hasegawa. "An Ethological and Emotional Basis for Human-Robot Interaction," *Robotics and Autonomous Systems,* vol.42, no. 3–4, pp.191–201, 2003.
- R. C. Arkin, "Moving up the Food Chain: Motivation and Emotion in Behavior-based Robots," in *Who Needs Emotions: The Brain Meets the Robot*, J. Fellous and M. Arbib., Eds., New York: Oxford University Press, 2005.

Affective Computing

- M. Boden, J. Bryson, D. Caldwell, et al. "Principles of Robotics: Regulating Robots in the Real World." *Connection Science,* vol. 29, no. 2, pp. 124–129, 2017.
- J. J Bryson, M. E. Diamantis and T. D. Grant. "Of, For, and By the People: The Legal Lacuna of Synthetic Persons," *Artificial Intelligence & Law,* vol. 25, no. 3, pp. 273–291, Sept. 2017.
- J. Novikova, and L. Watts, "Towards Artificial Emotions to Assist Social Coordination in HRI," *International Journal of Social Robotics,* vol. 7, no. 1, pp. 77–88, 2015.
- M. Scheutz, "The Affect Dilemma for Artificial Agents: Should We Develop Affective Artificial Agents?" *IEEE Transactions on Affective Computing,* vol. 3, no. 4, pp. 424–433, 2012.
- A. Sharkey and N. Sharkey. "Children, the Elderly, and Interactive Robots." *IEEE Robotics & Automation Magazine,* vol. 18, no. 1, pp. 32–38, 2011.

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Affective Computing Committee

- **Ronald C. Arkin** (Founding Co-Chair) – Regents' Professor & Director of the Mobile Robot Laboratory; College of Computing Georgia Institute of Technology
- **Joanna J. Bryson** (Co-Chair) – Reader (Associate Professor), University of Bath, Intelligent Systems Research Group, Department of Computer Science
- **John P. Sullins** (Co-Chair) – Professor of Philosophy, Chair of the Center for Ethics Law and Society (CELS), Sonoma State University
- **Genevieve Bell** – Intel Senior Fellow Vice President, Corporate Strategy Office Corporate Sensing and Insights
- **Jason Borenstein** – Director of Graduate Research Ethics Programs, School of Public Policy and Office of Graduate Studies, Georgia Institute of Technology
- **Cynthia Breazeal** – Associate Professor of Media Arts and Sciences, MIT Media Lab; Founder & Chief Scientist of Jibo, Inc.
- **Joost Broekens** – Assistant Professor Affective Computing, Interactive Intelligence group; Department of Intelligent Systems, Delft University of Technology
- **Rafael Calvo** – Professor & ARC Future Fellow, School of Electrical and Information Engineering, The University of Sydney
- **Laurence Devillers** – Professor of Computer Sciences, University Paris Sorbonne, LIMSI-CNRS 'Affective and social dimensions in spoken interactions' - member of the French Commission on the Ethics of Research in Digital Sciences and Technologies (CERNA)
- **Jonathan Gratch** – Research Professor of Computer Science and Psychology, Director for Virtual Human Research, USC Institute for Creative Technologie
- **Mark Halverson** – Founder and CEO at Human Ecology Holdings and Precision Autonomy
- **John C. Havens** – Executive Director, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems; Executive Director, The Council on Extended Intelligence; Author, *Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*

Affective Computing

- **Noreen Herzfeld** – Reuter Professor of Science and Religion, St. John's University
- **Chihyung Jeon** – Assistant Professor, Graduate School of Science and Technology Policy, Korea Advanced Institute of Science and Technology (KAIST)
- **Preeti Mohan** – Software Engineer at Microsoft and Computational Linguistics Master's Student at the University of Washington
- **Bjoern Niehaves** – Professor, Chair of Information Systems, University of Siegen
- **Rosalind Picard** – Rosalind Picard, (Sc.D, FIEEE) Professor, MIT Media Laboratory, Director of Affective Computing Research; Faculty Chair, MIT Mind+Hand+Heart; Co-founder & Chief Scientist, Empatica Inc.; Co-founder, Affectiva Inc.
- **Edson Prestes** – Professor, Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Brazil; Head, Phi Robotics Research Group, UFRGS; CNPq Fellow
- **Matthias Scheutz** – Professor, Bernard M. Gordon Senior Faculty Fellow, Tufts University School of Engineering
- **Robert Sparrow** – Professor, Monash University, Australian Research Council "Future Fellow", 2010-15.
- **Cherry Tom** – Emerging Technologies Intelligence Manager, IEEE Standards Association

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

Endnotes

[1] See B. J. Fogg, Persuasive technology. *Ubiquity*, December: 2, 2002.

[2] See S. Turkle, W. Taggart , C.D. Kidd, and O. Daste, "Relational artifacts with children and elders: the complexities of cybercompanionship, *Connection Science,* vol. 18, no. 4, 2006.

[3] A discussion of intimate robots for therapeutic and personal use is outside of the scope of *Ethically Aligned Design, First Edition*. For further treatment, among others, see J. P. Sullins, "Robots, Love, and Sex: The Ethics of Building a Love Machine." *IEEE Transactions on Affective Computing* 3, no. 4 (2012): 398–409.

[4] See R. Thaler, and C. R. Sunstein. *Nudge: Improving Decision about Health, Wealth and Happiness*, New Haven, CT: Yale University Press, 2008.

[5] See J. de Quintana Medina 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; and M. D. White, *The Manipulation of Choice. Ethics and Libertarian Paternalism.* New York: Palgrave Macmillan, 2013.

[6] See, for example, J. Borenstein and R. Arkin. "Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being." *Science and Engineering Ethics,* vol. 22, no. 1 (2016): 31–46.

[7] See S. Omohundro, "Autonomous Technology and the Greater Human Good." *Journal of Experimental and Theoretical Artificial Intelligence* 26, no. 3 (2014): 303–315.

[8] See J. Roughgarden, M. Oishi, and E. Akçay. "Reproductive Social Behavior: Cooperative Games to Replace Sexual Selection." *Science* 311, no. 5763 (2006): 965–969.

[9] See the Well-being chapter of this *Ethically Aligned Design, First Edition*.

Personal Data and Individual Agency

Regulations like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) of 2018 are helping to improve personal data protection. But legal compliance is not enough to mitigate the ethical implications and core challenges to human agency embodied by algorithmically driven behavioral tracking or persuasive computing. The core of the issue is one of parity.

Humans cannot respond on an individual basis to every algorithm tracking their behavior without technological tools supported by policy allowing them to do so. Individuals may provide consent without fully understanding specific terms and conditions agreements. But they are also not equipped to fully recognize how the nuanced use of their data to inform personalized algorithms affects their choices at the risk of eroding their agency.

Here we understand agency as an individual's ability to influence and shape their life trajectory as determined by their cultural and social contexts. Agency in the digital arena enables an individual to make informed decisions where their own terms and conditions can be recognized and honored at an algorithmic level.

To strengthen individual agency, governments and organizations must test and implement technologies and policies that let individuals create, curate, and control their online agency as associated with their identity. Data transactions should be moderated and case-by-case authorization decisions from the individual as to who can process what personal data for what purpose.

Specifically, we recommend governments and organizations:

- **Create:** Provide every individual with the means to create and project their own terms and conditions regarding their personal data that can be read and agreed to at a machine-readable level.
- **Curate:** Provide every individual with a personal data or algorithmic agent which they curate to represent their terms and conditions in any real, digital, or virtual environment.
- **Control**: Provide every individual access to services allowing them to create a trusted identity to control the safe, specific, and finite exchange of their data.

Three sections of this chapter reflect these core ideals regarding human agency.

A fourth section addresses issues surrounding personal data and individual agency relating to children.

Section 1—Create

To retain agency in the algorithmic era, each individual must have the means to create and project their own terms and conditions regarding their personal data. These must be readable and usable by both humans and machines.

Issue: What would it mean for a person to have individually controlled terms and conditions for their personal data?

Background

Part of providing individually controlled terms and conditions for personal data is to help each person consider what their preferences are regarding their data versus dictating how they need to share it. While questions along these lines are framed in light of a person's privacy, their preferences also reveal larger values for individuals. The ethical issue is whether A/IS act in accordance with these values.

This process of investigating one's values to identify these preferences is a powerful step towards regaining data agency. The point is not only that a person's data are protected, but also that by curating these answers they become educated about how important their information is in the context of how it is shared.

Most individuals also 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 in the future. Agreeing to most standard terms and conditions on these sites largely means users consent to give up control of their personal data rather than play a meaningful role in defining and curating its downstream use.

The scope of how long one should or could control the downstream use of their data can be difficult to calculate as consent-based models of personal data have trained users to release rights on any claims for use of their data which are entirely provided to the service, manufacturer, and their partners. However, models like YouTube's Content ID provide a form of precedent for thinking about how an individual's data could be technically protected where it is considered as an asset they could control and copyright. Here is language from YouTube's site about the service: "Copyright owners can use a system called Content ID to easily identify and manage their content on YouTube. Videos uploaded to YouTube are scanned against a database of files that have been submitted to us by content owners." In this sense, the question of how long or how far downstream one's personal data should be protected takes on the same logic of how long a corporation's intellectual property or copyrights could be protected based on initial legal terms set.

Personal Data and Individual Agency

One challenge is how to define use of data that can affect the individual directly, versus use of aggregated data. 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 data provided by a user could be included in an overall travel system's management database, aggregated into patterns for scheduling and maintenance as long as the individual-level data are deleted. Where users have predetermined via their terms and conditions that they are willing to share their data for these travel systems, they can meaningfully articulate how to share their information.

Under current business models, it is common for people to consent to the sharing of discrete data like credit card transaction data, answers to test questions, or how many steps they walk. However, once aggregated these data and the associated insights may lead to complex and sensitive conclusions being drawn about individuals. This end use of the individual's data may not have been part of the initial sharing agreement. This is why models for terms and conditions created for user control typically alert people via onscreen or other warning methods when their predetermined preferences are not being honored.

Recommendation

Individuals should be provided tools that produce machine-readable terms and conditions that are dynamic in nature and serve to protect their data and honor their preferences for its use.

Specifically:

- Personal data access and consent should be managed by the individual using their curated terms and conditions that provide notification and an opportunity for consent at the time data are exchanged, versus outside actors being able to access personal data without an individual's awareness or control.
- Terms should be presented in a way that allows a user to easily read, interpret, understand, and choose to engage with any A/IS. Consent should be both conditional and dynamic, where "dynamic" means downstream uses of a person's data must be explicitly called out, allowing them to cancel a service and potentially rescind or "kill" any data they have shared with a service to date via the use of a "Smart Contract" or specific conditions as described in mutual terms and conditions between two parties at the time of exchange.
- For further information on these issues, please see the following section in regard to algorithmic agents and their application.

Further Resources

- IEEE P7012™ - IEEE Standards Project for Machine Readable Personal Privacy Terms. This approved standardization project (currently in development) directly honors the goals laid out in Section One of this document.
- The Personalized Privacy Assistant Project Carnegie Mellon University. https://privacyassistant.org, 2019.

- M. Orcutt, "Personal AI Privacy Watchdog Could Help You Regain Control of Your Data" MIT Technology Review, May 11, 2017.
- M. Hintze, Privacy Statements: Purposes, Requirements, and Best Practices. Cambridge, U.K.: Cambridge University Press, 2017.
- D. J. Solove, "Privacy self-management and the consent dilemma, Harvard Law Review, vol. 126, no. 7, pp. 1880–1903, May 2013.
- N. Sadeh, M. Degeling, A. Das, A. S. Zhang, A. Acquisti, L. Bauer, L. Cranor, A. Datta, and D. Smullen, A Privacy Assistant for the Internet of Things: https://www.usenix.org/sites/default/files/soups17_poster_sadeh.pdf
- H. Lee, R. Chow, M. R. Haghighat, H. M. Patterson and A. Kobsa, "IoT Service Store: A Web-based System for Privacy-aware IoT Service Discovery and Interaction," *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops),* Athens, pp. 107-112, 2018.
- L. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall, and J. Reagle, "The Platform for Privacy Preferences 1.0 (P3P1.0) Specification," W3C Recommendation, [Online]. Available: www.w3.org/TR/P3P/, Apr. 2002.
- L. F. Cranor, "Personal Privacy Assistants in the Age of the Internet of Things," in World Economic Forum Annual Meeting, 2016.

Section 2—Curate

To retain agency in the algorithmic era, we must provide every individual with a personal data or algorithmic agent they curate to represent their terms and conditions in any real, digital, or virtual environment. This "agent" would be empowered to act as an individual's legal proxy in the digital and virtual arena. Oftentimes, the functionality of this agent will be automated, operating along the lines of current ad blockers which do not permit prespecified algorithms to access a user's data. For other situations that might be unique or new to this agent, a user could specify that notices or updates be sent on a case-by-case basis to determine where there could be a concern.

Issue: What would it mean for a person to have an algorithmic agent helping them actively represent and curate their terms and conditions at all times?

Background

While it's essential to create your own terms and conditions to broadcast your preferences, it's also important to recognize that humans do not operate at an algorithmic speed or level. A significant part of retaining your agency in this

Personal Data and Individual Agency

way involves identifying trusted services that can essentially act on your behalf when making decisions about your data.

Part of this logic entails putting you "at the center of your data". One of the greatest challenges to user agency is that once you give your data away, you do not know how it is being used or by whom. But when all transactions about your data go through your A/IS agent honoring your preferences, you have better opportunities to control how your information is shared.

As an example, with medical data—while it is assumed most would share all their medical data with their spouse—most would also not wish to share that same amount of data with their local gym. This is an issue that extends beyond privacy, meaning one's cultural or individual preferences about what personal information to share, to utility and clarity. This type of sharing also benefits users or organizations on the receiving end of data from these exchanges. For instance, the local gym in the previous example may only need basic heart or general health information and would actually not wish to handle or store sensitive cancer or other personal health data for reasons of liability.

A precedent for this type of patient- or user-centric model comes from Gliimpse, a service described by Jordan Crook from ***TechCrunch*** in his article, "Apple acquired Gliimpse, a personal health data startup": "Gliimpse works by letting users pull their own medical info into a single virtual space, with the ability to add documents and pictures to fill out the profile. From there, users can share that data (as a comprehensive picture) to whomever they wish." The fact that Apple acquired the startup points to the potential for the successful business model of user-centric data exchange and putting individuals at the center of their data.

A person's A/IS agent is a proactive algorithmic tool honoring their terms and conditions in the digital, virtual, and physical worlds. Any public space where a user may not be aware they are under surveillance by facial recognition, biometric, or other tools that could track, store, and utilize their data can now provide overt opportunity for consent via an A/IS agent platform. Even where an individual is not sure they are being tracked, by broadcasting their terms and conditions via digital means, they can demonstrate their preferences in the public arena. Via Bluetooth or similar technologies, individuals could offer their terms and conditions in a ubiquitous and always-on manner. This means even when an individual's terms and conditions are not honored, people would have the ability to demonstrate their desire not to be tracked which could provide a methodology for the democratic right to protest in a peaceful manner. And where those terms and conditions are recognized—meaning technically recognized even if they are not honored—one's opinions could be formally logged via GPS and timestamp data.

The A/IS agent 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 are being used in a way that was not authorized, or make recommendations to the user based on a personal profile. As a negotiator, the agent could broker conditions for sharing data and could include payment to the user as a

Personal Data and Individual Agency

term, or even retract consent for the use of data previously authorized, for instance, if a breach of conditions was detected.

Recommendations

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

- For purposes of privacy, a person must be able to set up complex permissions that reflect a variety of wishes.
- The agent should help a person foresee and mitigate potential ethical implications of specific machine learning data exchanges.
- A user should be able to override his/her personal agents should he/she decide that the service offered is worth the conditions imposed.
- An agent should enable machine-to-machine processing of information to compare, recommend, and assess offers and services.
- Institutional systems should ensure support for and respect the ability of individuals to bring their own agent to the relationship without constraints that would make some guardians inherently incompatible or subject to censorship.
- Vulnerable parts of the population will need protection in the process of granting access.

Further Resources

- IEEE P7006™ - IEEE Standards Project on Personal Data AI Agent Working Group. Designed as a tool to allow any individual to create their own personal "terms and conditions" for their data, the AI Agent will also provide a technological tool for individuals to manage and control their identity in the digital and virtual world.
- 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.

Section 3—Control

To retain agency in the algorithmic era, we must provide every individual access to services allowing them to create a trusted identity to control the safe, specific, and finite exchange of their data.

Issue: How can we increase agency by providing individuals access to services allowing them to create a trusted identity to control the safe, specific, and finite exchange of their data?

Background

Pervasive behavior-tracking adversely affects human agency by recognizing our identity in every action we take on and offline. This is why identity as it relates to individual data is emerging at the forefront of the risks and opportunities related to use of personal information for A/IS. Across the identity landscape there is increasing tension between the requirement for federated identities versus a range of identities. In federated identities, all data are linked to a natural and identified person. When one has a range of identities, or personas, these can be context specific and determined by the use case. 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, e.g., those issued by governments, banks, and regulatory authorities, to help put individuals at the center of their data in the algorithmic age.

Personas, identities that act as proxies, and pseudonymity are also critical requirements for privacy management and agency. These 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 that a user is over eighteen or is eligible for a service.

Attribute verification 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 of algorithms interacting with narrow segments of our identity data. In these situations, individuals typically are not aware of the context for how their data will be used.

Recommendation

Individuals should have access to trusted identity verification services to validate, prove, and support the context-specific use of their identity.

Further Resources

- Sovrin Foundation, The Inevitable Rise of Self-Sovereign Identity, Sept. 29, 2016.
- T. Ruff, "Three Models of Digital Identity Relationships," Evernym, Apr. 24, 2018.
- C. Pettey, The Beginner's Guide to Decentralized Identity. Gartner, 2018.
- C. Allen, The Path to Self-Sovereign Identity. GitHub, 2017.

Section 4—Children's Data Issues

While the focus of this chapter is to provide all individuals with agency regarding their personal data, some sectors of society have little or no control. For some elderly individuals or the mentally ill, it is because they have been found to not have "mental capacity", and for prisoners in the criminal justice system, society has taken control away as punishment. In the case of children, this is because they are considered human beings in development with evolving capacities.

We examine the issues of children as an example case and recommend either regulation or a technical architecture that provides a veil and buffer from harm until a child is at an age where they can claim personal responsibility for their decisions.

In many parts of the world, children are viewed by the law as being primarily charges of their parents who make choices on their behalf. In Europe, however, the state has a role in ensuring the "best interests of the child"[1] [2]. In schools, the two interests operate side-by-side, with parents being given some control over their child's education but with many decisions being made by the schools.

Many of the issues described above concern choices around personal data and the future impacts of how the data are gathered and shared. Children are at the forefront of technological developments with future educational and recreational technology gathering data from them all day at school and intelligent toys throughout their time at home.

As children post, click, search, and share information, their data are linked to various profiles, grouped into segmented audiences, and fed into machine learning algorithms. Some of these may be designed to target campaigns that increase sales, influence sentiment, encourage online games, impact social networks, or influence religious and political views. Data fed into algorithmic advertising is not only gathered from children's online actions but also from their devices. An example of device data is browser fingerprinting.[3] It includes a set of data about a child's browser or operating system. Fingerprinting vastly increases privacy risks because it is used to link to an individual.

Increasingly, children's beliefs and social norms are established by what they see and experience online. Their actions reflect what they believe is possible and expected. The report, "Digital Deceit: Technologies Behind Precision Propaganda on the Internet"[4], explains how companies collect, process, and then monetize personal preferences, socioeconomic status, fears, political and religious beliefs, location, and patterns of internet use.

Companies, governments, political parties, and philosophical and religious organizations use data available about students and children to influence how they spend their time, money, and the people or institutions they trust and with whom they spend time and build relationships.

Many aspects of a child's life can be digitized. Their behavioral, device, and network data are combined and used by machine learning

Personal Data and Individual Agency

algorithms to determine the information and content that best achieve the educational goals of the schools and the economic goals of the advertisers and platform companies.

Issue: Mass personalization of instruction

Background

The mass personalization of education offers better education for all at very low cost through A/IS-enabled computer-based instruction that promises to free up teachers to work with kids individually to pursue their passions. These applications will rely on the continuous gathering of personal data regarding mood, thought processes, private stories, physiological data, and more. The data will be used to construct a computational model of each child's interests, understanding, strengths, and weaknesses. The model provides an intimate understanding of how they think, what they understand, how they process information, or react to new information; all of which can be used to drive instructional content and feedback.

Sharing of this data between classes, enabling it to follow students through their schooling, will make the models more effective and beneficial to children, but it also exposes children and their families to social control. If performance data are correlated with social data on a family, it could be used by social authorities in decision-making about the family. For example, since 2015-2018, well-being digital tests were performed in schools in Denmark. Children were asked about everything from bullying, loneliness, and stomachaches. Recently it was disclosed that although the collected data was presented as anonymous, they were not. Data were stored with social security numbers, correlated with other test data, and even used in case management by some Danish municipalities.[5] Commercial profiling and correlation of different sets of personal data may further affect these children in future job or educational situations.

Recommendation

Educational data offer a unique opportunity to model individuals' thought processes and could be used to predict or change individuals' behavior in many situations. Governments and organizations should classify educational data as being sensitive and implement special protective standards.

Children's data should be held in "escrow" and not used for any commercial purposes until a child reaches the age of majority and is able to authorize use as they choose.

Further Resources

- The journal of the International Artificial Intelligence in Education Society: http://iaied.org/journal/
- Deeper discussion and bibliography of future trends of AI-based education with utopian and dystopian case scenarios: N. Pinkwart, "Another 25 Years of AIED? Challenges and Opportunities for Intelligent Educational Technologies of the Future," International Journal of Artificial Intelligence in Education, vol. 26, no. 2, pp. 771–783, 2016. [Online].

Personal Data and Individual Agency

Available: https://doi.org/10.1007/s40593-016-0099-7 [Accessed Dec. 2018].

- Information Commissioners Office (ico.),"What if we want to profile children or make automated decisions about them?" https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/children-and-the-gdpr/what-if-we-want-to-profile-children-or-make-automated-decisions-about-them/
- K. Firth-Butterfield, "What happens when your child's friend is an AI toy that talks back?" in World Economic Forum: Generation AI, https://www.weforum.org/agenda/2018/05/generation-ai-what-happens-when-your-childs-invisible-friend-is-an-ai-toy-that-talks-back/, May 22, 2018.

Issue: Technology choice-making in schools

Background

Children, as minors, have no standing to give or deny consent, or to control the use of their personal data. Parents only have limited choices in what are often school-wide implementations of educational technology. Examples include the use of Google applications, face recognition in security systems, and computer driven instruction as described above. In many cases, parents' only choice would be to send their children to a different school, but that choice is seldom available.

How should schools make these choices? How much input should parents have? Should parents be able to demand technology-free teaching?

There are many gaps in current student data regulation. In June 2018, CLIP, The Center on Law and Information Policy at Fordham Law School published, "Transparency and the Marketplace for Student Data".[6] This study concluded that "student lists are commercially available for purchase on the basis of ethnicity, affluence, religion, lifestyle, awkwardness, and even a perceived or predicted need for family planning services". Fordham found that the data market is becoming one of the largest and most profitable marketplaces in the United States. Data brokers have databases that store billions of data elements on nearly every United States consumer. However, information from students in the pursuit of an education should not be exploited and commercialized without restraint.

Fordham researchers found at least 14 data brokers who advertise the sale of student information. One sold lists of students as young as two years old. Another sold lists of student profiles on the basis of ethnicity, religion, economic factors, and even gawkiness.

Recommendation

Local and national educational authorities must work to develop policies surrounding students' personal data with all stakeholders: administrators, teachers, technology providers, students, and parents in order to balance the best educational interests of each child with the best practices to ensure safety of their personal data. Such efforts will raise awareness among all stakeholders of the promise and the compromises inherent in new educational technologies.

Personal Data and Individual Agency

Further Resources

- Common Sense Media privacy evaluation project: https://www.commonsense.org/education/privacy
- D. T. Ritvo, L. Plunkett, and P. Haduong,"Privacy and Student Data: Companion Learning Tools." Berkman Klein Center for Internet and Society at Harvard University, 2017. [Online]. Available: http://blogs.harvard.edu/youthandmediaalpha/files/2017/03/PrivacyStudentData_Companion_Learning_Tools.pdf [Accessed Dec. 2018].
- F. Alim, N. Cardozo, G. Gebhart, K. Gullo, and A. Kalia, "Spying on Students: School-Issued Devices and Student Privacy," Electronic Frontier Foundation, https://www.eff.org/wp/school-issued-devices-and-student-privacy, April 13, 2017.
- N. C. Russell, J. R. Reidenberg, E. Martin, and T. Norton, "Transparency and the Marketplace for Student Data," *Virginia Journal of Law and Technology*, Forthcoming. Available at SSRN: https://ssrn.com/abstract=3191436, June 6, 2018.

Issue: Intelligent toys

Background

Children will not only be exposed to A/IS at school but also at home, while they play and while they sleep. Toys are already being sold that offer interactive, intelligent opportunities for play. Many of them collect video and audio data which is stored on company servers and either is or could be mined for profiling or marketing data.

There is currently little regulatory oversight. In the United States COPPA[7] offers some protection for the data of children under 13. Germany has outlawed such toys using legislation banning spying equipment enacted in 1981. Corporate A/IS are being embodied in toys and given to children to play with, to talk to, tell stories to, and to explore all the personal development issues that we learn about in private play as children.

Recommendations

Child data should be held in "escrow" and not used for any commercial purposes until a child reaches the age of majority and is able to authorize use as they choose.

Governments and organizations need to educate and inform parents of the mechanisms of A/IS and data collection in toys and the possible impact on children in the future.

Further Resources

- K. Firth-Butterfield, "What happens when your child's friend is an AI toy that talks back?" in World Economic Forum: Generation AI, https://www.weforum.org/agenda/2018/05/generation-ai-what-happens-when-your-childs-invisible-friend-is-an-ai-toy-that-talks-back/, May 22, 2018.
- D. Basulto,"How artificial intelligence is moving from the lab to your kid's playroom," Washington Post, Oct. 15, 2015. [Online]. Available: https://www.washingtonpost.com/news/innovations/wp/2015/10/15/how-artificial-intelligence-is-moving-from-the-lab-to-your-kids-playroom/?utm_term=.89a1431a05a7 [Accessed Dec. 1, 2018].

- S. Chaudron, R. Di Gioia, M. Gemo, D. Holloway, J. Marsh, G. Mascheroni J. Peter, and D. Yamada-Rice , http://publications.jrc.ec.europa.eu/repository/handle/JRC105061, 2016.
- S. Chaudron, R. Di Gioia, M. Gemo, D. Holloway, J. Marsh, G. Mascheroni, J. Peter, D. Yamada-Rice Kaleidoscope on the Internet of Toys - Safety, security, privacy and societal insights, EUR 28397 EN, doi:10.2788/05383, Luxembourg: Publications Office of the European Union, 2017.
- Z. Kleinman, "Alexa, are you friends with our kids?" *BBC News,* July 16, 2018. [Online]. Available: https://www.bbc.com/news/technology-44847184.%5b. [Accessed Dec. 1, 2018].
- J. Wakefield, "Germany bans children's smartwatches." *BBC News,* Nov. 17 2017. [Online]. Available: https://www.bbc.co.uk/news/technology-42030109. [Accessed Dec. 2018].

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Personal Data and Individual Agency Committee

- **Katryna Dow** (Co-Chair) – CEO & Founder at Meeco
- **John C. Havens** (Co-Chair) – Executive Director, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems; Executive Director, The Council on Extended Intelligence; Author, *Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*
- **Mads Schaarup Andersen** – Senior Usable Security Expert in the Alexandra Institute's Security Lab
- **Ajay Bawa** – Technology Innovation Lead, Avanade Inc.
- **Ariel H. Brio** – Privacy and Data Counsel at Sony Interactive Entertainment
- **Walter Burrough** – Co-Founder, Augmented Choice; PhD Candidate (Computer Science) – Serious Games Institute
- **Danny W. Devriendt** – Managing director of Mediabrands Dynamic (IPG) in Brussels, and the CEO of the Eye of Horus, a global think-tank for communication-technology related topics
- **Dr. D. Michael Franklin** – Assistant Professor, Kennesaw State University, Marietta Campus, Marietta, GA
- **Jean-Gabriel Ganascia** – Professor, University Pierre et Marie Curie; LIP6 Laboratory ACASA Group Leader

Personal Data and Individual Agency

- **Bryant Joseph Gilot, MD CM DPhil MSc** – Center for Personalised Medicine, University of Tuebingen Medical Center, Germany & Chief Medical Officer, Blockchain Health Co., San Francisco
- **David Goldstein** – Seton Hall University
- **Adrian Gropper, M.D.** – CTO, Patient Privacy Rights Foundation; HIE of One Project
- **Marsali S. Hancock** – Chair, IEEE Standards for Child and Student Data governance, CEO and Co-Foundation EP3 Foundation.F
- **Gry Hasselbalch** – Founder DataEthics, Author, ***Data Ethics - The New Competitive Advantage***
- **Yanqing Hong** – Graduate, University of Utrecht Researcher at Tsinghua University
- **Professor Meg Leta Jones** – Assistant Professor in the Communication, Culture & Technology program at Georgetown University
- **Mahsa Kiani** – Chair of Student Activities, IEEE Canada; Vice Editor, IEEE Canada Newsletter (ICN); PhD Candidate, Faculty of Computer Science, University of New Brunswick
- **Brenda Leong** – Senior Counsel, Director of Operations, The Future of Privacy Forum
- **Emma Lindley** – Founder, Innovate Identity
- **Ewa Luger** – Chancellor's Fellow at the University of Edinburgh, within the Design Informatics Group
- **Sean Martin McDonald** – CEO of FrontlineSMS, Fellow at Stanford's Digital Civil Society Lab, Principal at Digital Public
- **Hiroshi Nakagawa** – Professor, The University of Tokyo, and AI in Society Research Group Director at RIKEN Center for Advanced Intelligence Project (AIP)
- **Sofia C. Olhede** – Professor of Statistics and an Honorary Professor of Computer Science at University College London, London, U.K; Member of the Programme Committee of the International Centre for Mathematical Sciences.
- **Ugo Pagallo** – University of Turin Law School; Center for Transnational Legal Studies, London; NEXA Center for Internet & Society, Politecnico of Turin
- **Dr. Juuso Parkkinen** – Senior Data Scientist, Nightingale Health; Programme Team Member, MyData 2017 conference
- **Eleonore Pauwels** – Research Fellow on AI and Emerging Cybertechnologies, United Nations University (NY) and Director of the AI Lab, Woodrow Wilson International Center for Scholars (DC)
- **Dr. Deborah C. Peel** – Founder, Patient Privacy Rights & Creator, the International Summits on the Future of Health Privacy
- **Walter Pienciak** – Principal Architect, Advanced Cognitive Architectures, Ltd.
- **Professor Serena Quattrocolo** – University of Turin Law School
- **Carolyn Robson** – Group Data Privacy Manager at Etihad Aviation Group
- **Gilad Rosner** – Internet of Things Privacy Forum; Horizon Digital Economy Research Institute, UK; UC Berkeley Information School

Personal Data and Individual Agency

- **Prof. Dr.-Ing. Ahmad-Reza Sadeghi** – Director System Security Lab, Technische Universität Darmstadt / Director Intel Collaborative Research Institute for Secure Computing
- **Rose Shuman** – Partner at BrightFront Group & Founder, Question Box
- **Dr. Zoltán Szlávik** – Lead/Researcher, IBM Center for Advanced Studies Benelux
- **Udbhav Tiwari** – Centre for Internet and Society, India

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

Endnotes

[1] Europäische Union, Europäischer Gerichtshof für Menschenrechte, & Europarat (Eds.). (2015). Handbook on European law relating to the rights of the child. Luxembourg: Publications Office of the European Union. https://www.echr.coe.int/Documents/Handbook_rights_child_ENG.PDF

[2] Children Act (1989). Retrieved from https://www.legislation.gov.uk/ukpga/1989/41/section/1

[3] "Browser fingerprints, and why they are so hard to erase | Network World." 17 Feb. 2015, https://www.networkworld.com/article/2884026/security0/browser-fingerprints-and-why-they-are-so-hard-to-erase.html. Accessed 25 July. 2018.

[4] D. Gosh and B. Scott, "Digital Deceit: The Technologies behind Precision Propaganda on the Internet" 23 Jan. 2018, https://www.newamerica.org/public-interest-technology/policy-papers/digitaldeceit/. Accessed 10 Nov 2018.

[5] Case described in Danish here https://dataethics.eu/trivsel-enhver-pris/

[6] Russell, N. Cameron, Reidenberg, Joel R., Martin, Elizabeth, and Norton, Thomas, "Transparency and the Marketplace for Student Data" (June 6, 2018). Virginia Journal of Law and Technology, Forthcoming. Available at SSRN: https://ssrn.com/abstract=3191436

[7] Children's Online Privacy Protection Act (COPPA) - https://www.ftc.gov/tips-advice/business-center/privacy-and-security/children%27s-privacy

Methods to Guide Ethical Research and Design

Autonomous and intelligent systems (A/IS) research and design must be developed against the backdrop that technology is not neutral. A/IS embody values and biases that can influence important social processes like voting, policing, and banking. To ensure that A/IS benefit humanity, A/IS research and design must be underpinned by ethical and legal norms. These should be instantiated through values-based research and design methods. Such methods put human well-being at the core of A/IS development.

To help achieve these goals, researchers, product developers, and technologists across all sectors need to embrace research and development methods that evaluate their processes, products, values, and design practices in light of the concerns and considerations raised in this chapter. This chapter is split up into three sections:

Section 1—Interdisciplinary Education and Research

Section 2—Corporate Practices on A/IS

Section 3—Responsibility and Assessment

Each of the sections highlights various areas of concern (issues) as well as recommendations and further resources.

Overall, we address both structural and individual approaches. We discuss how to improve the ethical research and business practices surrounding the development of A/IS and attend to the responsibility of the technology sector vis-à-vis the public interest. We also look at that what can be done at the level of educational institutions, among others, informing engineering students about ethics, social justice, and human rights. The values-based research and design method will require a change of current system development approaches for organizations. This includes a commitment of research institutions to strong ethical guidelines for research and of businesses to values that transcend narrow economic incentives.

Section 1—Interdisciplinary Education and Research

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

Issue: Integration of ethics in A/IS-related degree programs

Background

A/IS engineers and design teams do not always thoroughly explore the ethical considerations implicit in their technical work and design choices. Moreover, the overall science, technology, engineering, and mathematics (STEM) field struggles with the complexity of ethical considerations, which cannot be readily articulated and translated into the formal languages of mathematics and computer programming associated with algorithms and machine learning.

Ethical issues can easily be rendered invisible or inappropriately reduced and simplified in the context of technical practice. For the dangers of this approach see for instance, Lipton and Steinhardt (2018), listed under "Further Resources". This problem is further compounded by the fact that many STEM programs do not sufficiently integrate applied ethics throughout their curricula. When they do, often ethics is relegated to a stand-alone course or module that gives students little or no direct experience in ethical decision-making. Ethics education should be meaningful, applicable, and incorporate best practices from the broader field.

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

Recommendations

- Ethics training needs to be a core subject for all those in the STEM field, beginning at the earliest appropriate level and for all advanced degrees.
- Effective STEM ethics curricula should be informed by experts outside the STEM community from a variety of cultural and educational backgrounds to ensure that students acquire sensitivity to a diversity of robust perspectives on ethics and design.
- 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 and beyond should span primary, secondary, and postsecondary education, and include both universities and vocational training schools.

- Relevant accreditation bodies should reinforce this integrated approach as outlined above.

Further Resources

- IEEE P7000™ Standards Project for a Model Process for Addressing Ethical Concerns During System Design. IEEE P7000 aims to enhance corporate IT innovation practices by providing processes for embedding a values- and virtue-based thinking, culture, and practice into them.
- Z. Lipton and J. Steinhardt, Troubling Trends in Machine Learning Scholarship. ICML conference paper, July 2018.
- J. Holdren, and M. Smith. "Preparing for the Future of Artificial Intelligence." Washington, DC: Executive Office of the President, National Science and Technology Council, 2016.
- Comparing the UK, EU, and US approaches to AI and ethics: C. Cath, S. Wachter, B. Mittelstadt, et al., "Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach." *Science and Engineering Ethics*, vol. 24, pp. 505-528, 2017.

Issue: Interdisciplinary collaborations

Background

More institutional resources and incentive structures are necessary to bring 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 and shape the future of technological innovation. More could be done to develop methods, shared knowledge, and lexicons that would facilitate such collaboration.

This issue relates, among other things, to funding models as well as the lack of diversity of backgrounds and perspectives in A/IS-related institutions and companies, which limit cross-pollination between disciplines. To help bridge this gap, additional translation work and resource sharing, including websites and Massive Open Online Courses (MOOCs), need to happen among technologists and other relevant experts, e.g., in medicine, architecture, law, philosophy, psychology, and cognitive science. Furthermore, there is a need for more cross-disciplinary conversation and multi-disciplinary research, as is being done, for instance, at the annual ACM Fairness, Accountability, and Transparency (FAT*) conference or the work done by the Canadian Institute For Advanced Research (CIFAR), which is developing Canada's AI strategy.

Methods to Guide Ethical Research and Design

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

- S. Barocas, Course Material for Ethics and Policy in Data Science, Cornell University, 2017.
- L. Floridi, and M. Taddeo. "What Is Data Ethics?" *Philosophical Transactions of the Royal Society,* vol. 374, no. 2083, 1–4. DOI10.1098/rsta.2016.0360, 2016.
- S. Spiekermann, Ethical IT Innovation: A Value-Based System Design Approach. Boca Raton, FL: Auerbach Publications, 2015.
- K. Crawford, "Artificial Intelligence's White Guy Problem", *New York Times*, July 25, 2016. [Online]. Available: http://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=1. [Accessed October 28, 2018].

Issue: A/IS culture and context

Background

A responsible approach to embedding values into A/IS requires that algorithms and systems are created in a way that is sensitive to the variation of ethical practices and beliefs across cultures. The designers of A/IS need to be mindful of cross-cultural ethical variations while also respecting widely held international legal norms.

Recommendation

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 variation informs not only information flows and information systems, but also algorithmic decision-making and value by design.

Further Resources

- D. J. Pauleen, 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,* vol. 17, no. 17, 2006.
- J. Bielby, "Comparative Philosophies in Intercultural Information Ethics," *Confluence: Online Journal of World Philosophies* 2, no. 1, pp. 233–253, 2016.

Methods to Guide Ethical Research and Design

Issue: Institutional ethics committees 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 impact 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 the 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, necessary to govern ethical conduct of Human-Computer Interaction (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. Sections of IEEE standards governing research on A/IS in medical devices address some of the issues related to the security of A/IS-enabled devices. However, the ethics of testing those devices for the purpose of bringing them to market are not developed into policies or guidance documents from recognized national and international bodies, e.g., U.S. Food and Drug Administration (FDA) and EU European Medicines Agency (EMA). Second, the bodies that typically train individuals to be gatekeepers for the research ethics bodies are under-resourced in terms of expertise for A/IS development, e.g., Public Responsibility in Medicine and Research (PRIM&R) and the Society of Clinical Research Associates (SoCRA). 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.

For example, 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 the Institutional Animal Care and Use Committee (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.

Similarly, research pertinent to the intersection of radiological, biological, and toxicological research —ordinarily governed under institutional biosafety committees—and A/IS research is not often found in the literature pertinent to research ethics or research governance.

Methods to Guide Ethical Research and Design

Recommendation

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

Further Resources

- S. R. Jordan, "The Innovation Imperative." *Public Management Review* 16, no. 1, pp. 67–89, 2014.
- B. Schneiderman, "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, 13538–13540, 2016.
- J. Metcalf and K. Crawford, "Where are Human Subjects in Big Data Research? The Emerging Ethics Divide." *Big Data & Society,* May 14, 2016. [Online]. Available: SSRN: https://ssrn.com/abstract=2779647. [Accessed Nov. 1, 2018].
- R. Calo, "Consumer Subject Review Boards: A Thought Experiment." *Stanford Law Review Online* 66 97, Sept. 2013.

Section 2—Corporate Practices on A/IS

Corporations 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, and use.

Issue: Values-based ethical culture and practices for industry

Background

Corporations are built to create profit while competing for market share. This can lead corporations to focus on growth at the expense of avoiding negative ethical consequences. Given the deep ethical implications of widespread deployment of A/IS, in addition to laws and regulations, there is a need to create values-based ethical culture and practices for the development and deployment of those systems. To do so, we need to further identify and refine corporate processes that facilitate values-based design.

Recommendations

The building blocks of such practices include top-down leadership, bottom-up empowerment, ownership, and responsibility, along with the need to consider system deployment contexts and/or ecosystems. Corporations should identify stages in their processes in which ethical considerations, "ethics filters", are in place before products are further developed and deployed.

For instance, if an ethics review board comes in at the right time during the A/IS creation process, it would help mitigate the likelihood of creating ethically problematic designs. 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

- ACM Code of Ethics and Professional Ethics, which includes various references to human well-being and human rights, 2018.
- Report of UN Special Rapporteur on Freedom of Expression. *AI and Freedom of Expression.* 2018.
- The website of the Benefit corporations (B-corporations) provides a good overview of a range of companies that personify this type of culture.
- R. Sisodia, J. N. Sheth and D. Wolfe, Firms of Endearment, 2nd edition. Upper Saddle River, NJ: FT Press, 2014. This book showcases how companies embracing values and a stakeholder approach outperform their competitors in the long run.

Methods to Guide Ethical Research and Design

Issue: Values-based 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 A/IS. Cultivating an ethical corporate culture is an essential component of successful leadership in the A/IS domain.

Recommendations

Companies should create roles for senior-level marketers, engineers, and lawyers who can collectively and pragmatically implement ethically aligned design. There is also a need for more in-house ethicists, or positions that fulfill similar roles. One potential way to ensure values are on the agenda in A/IS development is to have a Chief Values Officer (CVO), a role first suggested by Kay Firth-Butterfield, see "Further Resources". However, ethical responsibility should not be delegated solely to CVOs. They can support the creation of ethical knowledge in companies, but in the end, all members of an organization will need to act responsibly throughout the design process.

Companies need to ensure that their understanding of values-based system innovation is based on *de jure* and *de facto* international human rights standards.

Further Resources

- K. Firth-Butterfield, "How IEEE Aims to Instill Ethics in Artificial Intelligence Design," The Institute. Jan. 19, 2017. [Online]. Available: http://theinstitute.ieee.org/ieee-roundup/blogs/blog/how-ieee-aims-to-instill-ethics-in-artificial-intelligence-design. [Accessed October 28, 2018].
- 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, ICT Sector Guide on Implementing the UN Guiding Principles on Business and Human Rights, 2013.
- C. Cath, and L. Floridi, "The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights." *Science and Engineering Ethics,* vol. 23, no. 2, pp. 449–468, Apr. 2017.

Issue: Empowerment to raise ethical concerns

Background

Engineers and design teams may 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, including the design, development, and deployment

Methods to Guide Ethical Research and Design

phases. Because raising ethical concerns can be perceived as slowing or halting a design project, organizations need to consider how they can recognize and incentivize values-based design as an integral component of product development.

Recommendations

Employees should be empowered and encouraged to raise ethical concerns in day-to-day professional practice.

To be effective in ensuring adoption of ethical considerations during product development or internal implementation of A/IS, organizations should create a company culture and set of norms that encourage incorporating ethical considerations in the design and implementation processes.

New categories of considerations around these issues need to be accommodated, along with updated Codes of Conduct, company value-statements, and other management principles so individuals are empowered to share their insights and concerns in an atmosphere of trust. Additionally, bottom-up approaches like company "town hall meetings" should be explored that reward, rather than punish, those who bring up ethical concerns.

Further Resources

- The British Computer Society (BCS), Code of Conduct, 2019.
- C. Cath, and L. Floridi, "The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights," *Science and Engineering Ethics,* vol. 23, no. 2, pp. 449–468, Apr. 2017.

Issue: Ownership and responsibility

Background

There is variance within the technology community on how it sees its responsibility regarding A/IS. The difference in values and behaviors are not necessarily aligned with the broader set of social concerns raised by public, legal, and professional communities. The current makeup of most organizations has clear delineations among engineering, legal, and marketing functions. Thus, technologists will often be incentivized in terms of meeting functional requirements, deadline, and financial constraints, 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 behavior 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. However, ethics in A/IS design include broader considerations about the consequences of technologies.

Recommendations

Organizations should clarify the relationship between professional ethics and applied A/IS ethics by helping or enabling designers, engineers, and other company representatives to discern the differences between these kinds of ethics and where they complement each other.

Methods to Guide Ethical Research and Design

Corporate ethical review boards, or comparable mechanisms, should be formed to address ethical and behavioral concerns in relation to A/IS design, development and deployment. 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.

Further Resources

- HH van der Kloot Meijberg and RHJ ter Meulen, "Developing Standards for Institutional Ethics Committees: Lessons from the Netherlands," *Journal of Medical Ethics* 27 i36-i40, 2001.

Issue: Stakeholder inclusion

Background

The interface between A/IS and practitioners, as well as other stakeholders, is gaining broader attention in domains such as healthcare 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 might be the "end user" of a robot or social A/IS technology. In order to develop a product that is ethically aligned, stakeholders' feedback is crucial to design a system that takes ethical and social issues into account. There are successful user experience (UX) design concepts, such as accessibility, that consider human physical disabilities, which should be incorporated into A/IS as they are more widely deployed. It is important to continuously consider the impact of A/IS through unanticipated use and on unforeseen interests.

Recommendations

To ensure representation of stakeholders, organizations should enact a planned and controlled set of activities to account for the interests of the full range of stakeholders or practitioners who will be working alongside A/IS and incorporating their insights to build upon, rather than circumvent or ignore, the social and practical wisdom of involved practitioners and other stakeholders.

Further Resources

- C. Schroeter, 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. pp. 1145–1151.
- T. L. Chen, et al. "Robots for Humanity: Using Assistive Robotics to Empower People with Disabilities," *IEEE Robotics and Automation Magazine,* vol. 20, no. 1, pp. 30–39, 2013.
- R. Hartson, and P. S. Pyla. *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Waltham, MA: Elsevier, 2012.

Methods to Guide Ethical Research and Design

Issue: Values-based design

Background

Ethics are often treated as an impediment to innovation, even among those who ostensibly support ethical design practices. In industries that reward rapid innovation in particular, it is necessary to develop ethical design practices that integrate effectively with existing engineering workflows. Those who advocate for ethical design within a company should be seen as innovators seeking the best 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 life cycles.

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. The transition points 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.

Recommendations

Companies should study design processes to identify situations where engineers and researchers can be encouraged to raise and resolve questions of ethics and foster a proactive environment to realize ethically aligned design. Achieving a distributed responsibility for ethics requires that all people involved in product design are encouraged to notice and respond to ethical concerns. Organizations should consider how they can best encourage and facilitate deliberations among peers.

Organizations should identify points for formal review during product development. 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

- A. Sinclair, "Approaches to Organizational Culture and Ethics," *Journal of Business Ethics,* vol. 12, no. 1, pp. 63–73, 1993.
- Al Y. S. Chen, R. B. Sawyers, and P. F. Williams. "Reinforcing Ethical Decision Making Through Corporate Culture," *Journal of Business Ethics* 16, no. 8, pp. 855–865, 1997.
- K. Crawford and R. Calo, "There Is a Blind Spot in AI Research," *Nature* 538, pp. 311–313, 2016.

Section 3—Responsibility and Assessment

Lack of accountability of the A/IS design and development process presents a challenge to ethical implementation and oversight. This section presents four issues, moving from macro oversight to micro documentation practices.

Issue: Oversight for algorithms

The algorithms behind A/IS are not subject to consistent oversight. This lack of assessment causes concern because end users have no account of how a certain algorithm or system came to its conclusions. These recommendations are similar to those made in the "General Principles" and "Embedding Values into Autonomous and Intelligent Systems" chapters of *Ethically Aligned Design*, but here the recommendations are used as they apply to the narrow scope of this chapter .

Recommendations

Accountability: As touched on in the General Principles chapter 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, or will not, be released to the general public. Nonetheless, standards providing oversight of the manufacturing process of A/IS technologies need to be created to avoid harm and negative consequences. 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.

Human rights and algorithmic impact assessments should be explored as a meaningful way to improve the accountability of A/IS. These need to be paired with public consultations, and the final impact assessments must be made public.

Further Resources

- F. Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press, 2016.
- R. Calo, "Artificial Intelligence Policy: A Primer and Roadmap," *UC Davis Law Review,* 52: pp. 399–435, 2017.
- ARTICLE 19. "Privacy and Freedom of Expression in the Age of Artificial Intelligence," Privacy International, April 2018. [Online]. Available: https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf. [Accessed October 28, 2018].

Methods to Guide Ethical Research and Design

Issue: Independent review organization

Background

We need independent, expert opinions that provide guidance to the general public regarding A/IS. 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 which 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 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.

Recommendations

An independent, internationally coordinated body—akin to ISO—should be formed to oversee whether A/IS products actually meet ethical criteria, both when designed, developed, deployed, and when considering their evolution after deployment and during interaction with other products. It should also include a certification process.

Further Resources

- A. Tutt, "An FDA for Algorithms," *Administrative Law Review* 69, 83–123, 2016.
- M. U. Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," *Harvard Journal of Law and Technology* vol. 29, no. 2, 354–400, 2016.
- D. R. Desai and J. A. Kroll, "Trust But Verify: A Guide to Algorithms and the Law." *Harvard Journal of Law and Technology*, Forthcoming; Georgia Tech Scheller College of Business Research Paper No. 17-19, 2017.

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 and intelligent systems, are a growing source of black box software. At least for the foreseeable future, A/IS developers will likely be unable to build systems that are guaranteed to operate as intended.

Methods to Guide Ethical Research and Design

Recommendations

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.

Technologists should be able to characterize what their algorithms or systems are going to do via documentation, audits, and 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. As such, it is also important to ensure access to remedy adverse impacts.

Technologists and corporations must do their ethical due diligence before deploying A/IS technology. Standards for what constitutes ethical due diligence would ideally be generated by an international body such as IEEE or ISO, and barring that, each corporation should work to generate a set of ethical standards by which their processes are evaluated and modified. Similar to a flight data recorder in the field of aviation, algorithmic traceability can provide insights on what computations led to questionable or dangerous behaviors. Even where such processes remain somewhat opaque, technologists should seek indirect means of validating results and detecting harms.

Further Resources

- M. Ananny and K. Crawford, "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability," *New Media & Society*, vol. 20, no. 3, pp. 973-989, Dec. 13, 2016.
- D. Reisman, J. Schultz, K. Crawford, and M. Whittaker, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability," AI NOW 2018. [Online]. Available: https://ainowinstitute.org/aiareport2018.pdf. [Accessed October 28, 2018].
- J. A. Kroll "The Fallacy of Inscrutability." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, C. Cath, S. Wachter, B. Mittelstadt and L. Floridi, Eds., October 15, 2018 DOI: 10.1098/rsta.2018.0084.

Issue: Need for better technical documentation

Background

A/IS are often construed as fundamentally opaque and inscrutable. However, lack of transparency is often the result of human decision. The problem can be traced to a variety of sources, including poor documentation that excludes vital information about the limitations and assumptions of a system. Better documentation combined with internal and external auditing are crucial to understanding a system's ethical impact.

Methods to Guide Ethical Research and Design

Recommendation

Engineers should be required to thoroughly document the end product and related data flows, performance, limitations, and risks of A/IS. Behaviors and practices 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 training data used in predictive systems, algorithms and components used, and results of behavior monitoring. Criteria for such documentation could be: auditability, accessibility, meaningfulness, and readability.

Companies should make their systems auditable and should explore novel methods for external and internal auditing.

Further Resources

- S. Wachter, B. Mittelstadt, and L. Floridi. "Transparent, Explainable, and Accountable AI for Robotics." *Science Robotics, vol.* 2, no. 6, May 31, 2017. [Online]. Available: DOI: 10.1126/scirobotics.aan6080. [Accessed Nov.
- S. Barocas, and A. D. Selbst, "Big Data's Disparate Impact." *California Law Review* 104, 671-732, 2016.
- J. A. Kroll, 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, 633–705, 2017.
- J. M. Balkin, "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation." *UC Davis Law Review*, 2017.

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Methods to Guide Ethical Research and Design Committee

- **Corinne Cath-Speth** (Co-Chair) – PhD student at Oxford Internet Institute, The University of Oxford, Doctoral student at the Alan Turing Institute, Digital Consultant at ARTICLE 19
- **Raja Chatila** (Co-Chair) – CNRS-Sorbonne Institute of Intelligent Systems and Robotics, Paris, France; Member of the French Commission on the Ethics of Digital Sciences and Technologies CERNA; Past President of IEEE Robotics and Automation Society
- **Thomas Arnold** – Research Associate at Tufts University Human-Robot Interaction Laboratory
- **Jared Bielby** – President, Netizen Consulting Ltd; Chair, International Center for Information Ethics; editor, *Information Cultures in the Digital Age*
- **Reid Blackman, PhD** – Founder & CEO Virtue Consultants, Assistant Professor of Philosophy Colgate University
- **Tom Guarriello, PhD** – Founding Faculty member in the Master's in Branding program at New York City's School of Visual Arts, Host of RoboPsyc Podcast and author of RoboPsych Newsletter
- **John C. Havens** – Executive Director, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems; Executive Director, The Council on Extended Intelligence; Author, *Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*
- **Sara Jordan** – Assistant Professor of Public Administration in the Center for Public Administration & Policy at Virginia Tech
- **Jason Millar** – Professor, robot ethics at Carleton University
- **Sarah Spiekermann** – Chair of the Institute for Information Systems & Society at Vienna University of Economics and Business; Author of the textbook "Ethical IT-Innovation", the popular book "Digitale Ethik—Ein Wertesystem für das 21. Jahrhundert" and Blogger on "The Ethical Machine"
- **Shannon Vallor** – William J. Rewak Professor in the Department of Philosophy at Santa Clara University in Silicon Valley and Executive Board member of the Foundation for Responsible Robotics
- **Klein, Wilhelm E. J., PhD** – Senior Research Associate & Lecturer in Technology Ethics, City University of Hong Kong

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

A/IS for Sustainable Development

Autonomous and intelligent systems (A/IS) offer unique and impactful opportunities as well as risks both to people living in high-income countries (HIC) and in low-and middle-income countries (LMIC). The scaling and use of A/IS represent a genuine opportunity across the globe to provide individuals and communities—be they rural, semi-urban, or urban—with the means to satisfy their needs and develop their full potential, with greater autonomy and choice. A/IS will potentially disrupt economic, social, and political relationships and interactions at many levels. Those disruptions could provide an historical opportunity to reset those relationships in order to distribute power and wealth more equitably and thus promote social justice.[1] They could also leverage quality and better standards of life and protect people's dignity, while maintaining cultural diversity and protecting the environment.

One possible vehicle that can be used to agree on priorities and prioritize resources and actions is the United Nations Agenda for Sustainable Development, which was adopted by the UN General Assembly in 2015; 193 nations voted in favor of the Agenda, which also includes 17 Sustainable Development Goals (SDGs) for the world to achieve by 2030. The Agenda challenges all member states to make concerted efforts toward the above mentioned goals, and thus toward a sustainable, prosperous, and resilient future for people and the planet. These universally applicable goals should be reached by 2030.[2]

The value of A/IS is significantly associated with the generation of various types of superior and unique insights, many of which could help achieve positive socioeconomic outcomes for both HIC and LMIC societies, in keeping with the SDGs. The ethical imperative driving this chapter is that A/IS must be harnessed to benefit humanity, promote equality, and realize the world community's vision of a sustainable future and the SDGs:

…….of universal respect for human rights and human dignity, the rule of law, justice, equality and nondiscrimination; of respect for race, ethnicity and cultural diversity; and of equal opportunity permitting the full realization of human potential and contributing to shared prosperity. A world which invests in its children and in which every child grows up free from violence and exploitation. A world in which every woman and girl enjoys full gender equality and all legal, social and economic barriers to their empowerment have been removed. A just, equitable, tolerant, open and socially inclusive world in which the needs of the most vulnerable are met.[3]

A/IS for Sustainable Development

We recognize that how A/IS are deployed globally will be a determining factor in whether, in fact, "no-one gets left behind", whether human rights and dignity of all people are respected, whether children are protected, and whether the gap between rich and poor, within and between nations, narrows or widens. A/IS can advance the Sustainable Development Agenda's transformative vision, but at the same time, A/IS can undermine it if risks reviewed in this chapter are not managed properly.

For example, A/IS create the risk of accelerating inequality within and among nations, if their development and marketing are controlled by a few select companies, primarily in HIC. The benefits would largely accrue to the highly educated and wealthier segment of the population, while displacing the less educated workforce, both by automation and by the absence of educational or retraining systems capable of imparting skills and knowledge needed to work productively alongside A/IS. These risks, although differentiated by IT infrastructure, educational attainment, economic, and cultural contexts, exist in HIC and LMIC alike. The inequality in accessing and using the internet, both within and among countries, raises questions on how to spread A/IS benefits across humanity. Ensuring A/IS "for the common good" is an ethical imperative and at the core of *Ethically Aligned Design, First Edition*; the key elements of this "common good" are that it is human-centered, accountable, and ensure outcomes that are fair and inclusive.

This chapter explores the imperative for A/IS to serve humanity by improving the quality and standard of life for all people everywhere. It makes recommendations for advancing equal access to this transformative technology, so that it drives the well-being of all people, rather than further concentrating wealth, resources, and decision-making power in the hands of a few countries, companies, or citizens. The recommendations further reflect policies and collaborative public, private, and people programs which, if implemented, will respect the ethical imperative embedded in the Sustainable Development Agenda's transformative vision. The respect of human rights and dignity, and the advancement of "common good" with equal benefit to both HIC and LMIC, are central to every recommendation within this chapter.

Section 1—A/IS in Service to Sustainable Development for All

A/IS have the potential to contribute to the resolution of some of the world's most pressing problems, including: violation of fundamental rights, poverty, exploitation, climate change, lack of high-quality services to excluded populations, increased violence, and the achievement of the SDGs.

Issue: Current roadmaps for development and deployment of A/IS are not aligned with or guided by their impact in the most important challenges of humanity, defined in the seventeen United Nations Sustainable Development Goals (SDGs), which collectively aspire to create a more equal world of prosperity, peace, planet protection, and human dignity for all people.[4]

Background

SDGs promoting prosperity, peace, planet protection, human dignity, and respect for human rights of all, apply to HIC and LMIC alike. Yet ensuring that the benefits of A/IS will accrue to humanity as a whole, leaving "no one behind", requires an ethical commitment to global citizenship and well-being, and a conscious effort to counter the nature of the tech economy, with its tendency to concentrate wealth within high income populations. Implementation of the SDGs should benefit excluded sectors of society in every country, regardless of A/IS infrastructure.

"The Road to Dignity by 2030" document of the UN Secretary General reports on resources and methods for implementing the 2030 Agenda for Sustainable Development and emphasizes the importance of science, technology, and innovation for a sustainable future.[5] The UN Secretary General posits that:

"A sustainable future will require that we act now to phase out unsustainable technologies and to invest in innovation and in the development of clean and sound technologies for sustainable development. We must ensure that they are fairly priced, broadly disseminated and fairly absorbed, including to and by developing countries." (para. 120)

A/IS are among the technologies that can play an important role in the solution of the deep social problems plaguing our global civilization, contributing to the transformation of society away from an unsustainable, unequal socioeconomic system, towards one that realizes the vision of universal human dignity, peace, and prosperity.

However, with all the potential benefits of A/IS, there are also risks. For example, given A/IS technology's immense power needs, without

A/IS for Sustainable Development

new sources of sustainable energy harnessed to power A/IS in the future, there is a risk that it will increase fossil fuel use and have a negative impact on the environment and the climate.

While 45% of the world's population is not connected to the internet, they are not necessarily excluded from A/IS' potential benefits: in LMIC mobile networks can provide data for A/IS applications. However, only those connected are likely to benefit from the income-producing potential of internet technologies. In 2017, internet penetration in HIC left behind certain portions of the population often in rural or remote areas; 12% of U.S. residents and 20% of residents across Europe were unable to access the internet. In Asia with its concentration of LMIC, 52% of the population, on average, had no access, a statistic skewed by the large population of China, where internet penetration reached 45% of the population. In numerous other countries in the region, 99% of residents had no access. This nearly total exclusion also exists in several countries in Africa, where the overall internet penetration is only 35%: 2 of every 3 residents in Africa have no access.[6] Those with no internet access also do not generate data needed to "train" A/IS, and are thereby excluded from benefits of the technology, the development of which risks systematic discriminatory bias, particularly against people from minority populations, and those living in rural areas, or in low-income countries. As a comparison, one study estimated that "in the US, just one home automation product can generate a data point every six seconds."[7] In Mozambique, where about 90% of the population lack internet access, "the average household generates zero digital data points."[8] With mobile phones generating much of the data needed for developing A/IS applications in LMIC, unequal phone ownership may build in bias. For example, there is a risk of discrimination against women, who across LMIC are 14% less likely than men to own a mobile phone, and in South Asia where 38% are less likely to own a mobile phone.[9]

Recommendations

The current range of A/IS applications in sectors crucial to the SDGs, and to excluded populations everywhere, should be studied, with the strengths, weaknesses, and potential of the most significant recent applications analyzed, and the best ones developed at scale. Specific objectives to consider include:

- Identifying and experimenting with A/IS technologies relevant to the SDGs, such as: big data for development relevant to, for example, agriculture and medical tele-diagnosis; geographic information systems needed in public service planning, disaster prevention, emergency planning, and disease monitoring; control systems used in, for example, naturalizing intelligent cities through energy and traffic control and management of urban agriculture; applications that promote human empathy focused on diminishing violence and exclusion and increasing well-being.
- Promoting the potential role of A/IS in sustainable development by collaboration between national and international government agencies and nongovernmental organizations (NGOs) in technology sectors.

A/IS for Sustainable Development

- Analyzing the cost of and proposing strategies for publicly providing internet access for all, as a means of diminishing the gap in A/IS' potential benefit to humanity, particularly between urban and rural populations in HIC and LMIC alike.
- Investing in the documentation and dissemination of innovative applications of A/IS that advance the resolution of identified societal issues and the SDGs.
- Researching sustainable energy to power A/IS computational capacity.
- Investing in the development of transparent monitoring frameworks to track the concrete results of donations by international organizations, corporations, independent agencies, and the State, to ensure efficiency and accountability in applied A/IS.
- Developing national legal, policy, and fiscal measures to encourage competition in the A/IS domestic markets and the flourishing of scalable A/IS applications.
- Integrating the SDGs into the core of private sector business strategies and adding SDG indicators to companies' key performance indicators, going beyond corporate social responsibility (CSR).
- Applying the well-being indicators[10] to evaluate A/IS' impact from multiple perspectives in HIC and LMIC alike.

Further Resources

- R. Van Est and J.B.A. Gerritsen, with assistance of L. Kool, Human Rights in the Robot Age: Challenges arising from the use of Robots, Artificial Intelligence and Augmented Reality Expert Report written for the Committee on Culture, Science, Education and Media of the Parliamentary Assembly of the Council of Europe (PACE), The Hague: Rathenau Instituut 2017.
- World Economic Forum Global Future Council on Human Rights 2016-18, "White Paper: How to Prevent Discriminatory Outcomes in Machine Learning," World Economic Forum, March 2018.
- United Nations General Assembly, *Transforming Our World: The 2030 Agenda for Sustainable Development* (A/RES/70/1: 21 October 2015) Preamble. http://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_70_1_E.pdf.
- United Nations Global Pulse, Big Data for Development: Challenges and Opportunities, 2012.

A/IS for Sustainable Development

Issue: A/IS are often viewed only as having impact in market contexts, yet these technologies also have an impact on social relations and culture.

Background

A/IS are expected to have an impact beyond market domains and business models, diffusing throughout the global society. For instance, A/IS have and will impact social relationships in a way similar to how mobile phones changed our daily lives, reflecting directly on our culture, customs, and language. The extent and direction of this impact is not yet clear, but documented experience in HIC and high internet-penetration environments of trolls, "fake news," and cyberbullying on social media offer a cautionary tale.[11] Depression, social isolation, aggression, and the dissemination of violent behavior with damage to human relations, so extreme that, in some cases, it has resulted in suicide, are all correlated with the internet.[12] As an example, the technology for "smart homes" has been used for inflicting domestic violence by remotely locking doors, turning off heat/AC, and otherwise harassing a partner. This problem could be easily extended to include elder and child abuse.[13] Measures need to be developed to prevent A/IS from contributing to the emergence or amplification of social disorders.

Recommendations

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

- Exploring the development of algorithms capable of detecting and reporting discrimination, cyberbullying, deceptive content and identities, etc., and of notifying competent authorities; recognizing that the use of such algorithms must address ethical concerns related to algorithm explainability as well as take into account the risk to certain aspects of human rights, notably to privacy and freedom from oppression.
- Developing a globally recognized professional Code of Ethics with and for technology companies.
- Identifying social disorders, such as depression, anxiety, psychological violence, political manipulation, etc., correlated with the use of A/IS-based technologies as a world health problem; monitoring and measuring their impact.
- Elaborating metrics measuring how, where and on whom there is a cultural impact of new A/IS-based technologies.

A/IS for Sustainable Development

Further Resources

- T. Luong, "Thermostats, Locks and Lights: Digital Tools of Domestic Abuse," *The New York Times*, June 23, 2018, https://www.nytimes.com/2018/06/23/technology/smart-home-devices-domestic-abuse.html.
- J. Naughton, "The internet of things has opened up a new frontier of domestic abuse." The Guardian, July 2018.
- M. Pianta, Innovation and Employment, Handbook of Innovation. Oxford, U.K.: Oxford University Press, 2003.
- M.J. Salganik, Bit by Bit. Princeton, NJ: Princeton University Press 2018.
- J. Torresen, "A Review of Future and Ethical Perspectives of Robotics and AI" Frontiers in Robotics and AI, Jan. 15, 2018. [Online]. Available: https://doi.org/10.3389/frobt.2017.00075. [Accessed Nov. 1, 2018].

Issue: The right to truthful information is key to a democratic society and to achieving sustainable development and a more equal world, but A/IS poses risks to this right that must be managed.

Background

Social media have become the dominant technological infrastructure for the dissemination of information such as news, opinion, advertising, etc., and are currently in the vanguard of the movement toward customized/targeted information based on user profiling that involves significant use of A/IS techniques. Analysis of opinion polls and trends in social networks, blogs, etc., and of the emotional response to news items can be used for the purposes of manipulation, facilitating both the selection of news that guides public opinion in the desired direction and the practice of sensationalism.

The "personalization of the consumer experience", that is, the adaptation of articles to the interests, political vision, cultural level, education, and geographic location of the reader, is a new challenge for the journalism profession that expands the possibilities of manipulation.

The information infrastructure is currently lacking in transparency, such that it is difficult or impossible to know (except perhaps for the infrastructure operator):

- what private information is being collected for user profiling and by whom,
- which groups are targeted and by whom,
- what information has been received by any given targeted group,
- who financed the creation and dissemination of this information,
- the percentage of the information being disseminated by bots, and
- who is financing these bots.

Many actors have found this opaque infrastructure ideal for spreading politically motivated disinformation, which has a negative

A/IS for Sustainable Development

effect on the creation of a more equal world, democracy, and the respect for fundamental rights. This disinformation can have tragic consequences. For instance, human rights groups have unearthed evidence that the military authorities of Myanmar used Facebook for inciting hatred against the Rohingya Muslim minority, hatred which facilitated an ethnic cleansing campaign and the murder of up to 50,000 people.[14] The UN determined that these actions constituted genocide, crimes against humanity, and war crimes.[15]

Recommendations

To protect democracy, respect fundamental rights, and promote sustainable development, governments should implement a legislative agenda which prevents the spread of misinformation and hate speech, by:

- Ensuring more control and transparency in the use of A/IS techniques for user profiling in order to protect privacy and prevent user manipulation.
- Using A/IS techniques to detect untruthful information circulating in the infrastructures, overseen by a democratic body to prevent potential censorship.
- Obliging companies owning A/IS infrastructures to provide more transparency regarding their algorithms, sources of funding, services, and clients.
- Defining a new legal status somewhere between "platforms" and "content providers" for A/IS infrastructures.
- Reformulating the deontological codes of the journalistic profession to take into account the intensive use of A/IS techniques foreseen in the future.
- Promoting the right to information in official documents, and developing A/IS techniques to automate journalistic tasks such as verification of sources and checking the accuracy of the information in official documents, or in the selection, hierarchy, assessment, and development of news, thereby contributing to objectivity and reliability.

Further Resources

- M. Broussard, "Artificial Iintelligence for Investigative Reporting: Using an expert system to enhance journalists' ability to discover original public affairs stories." Digital Journalism, vol. 3, no. 6, pp. 814-831, 2015.
- M. Carlson, "The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority." Digital Journalism, vol. 3, no. 3, pp. 416-431, 2015.
- A. López Barriuso, F. de la Prieta Pintado, Á. Lozano Murciego, , D. Hernández de la Iglesia and J. Revuelta Herrero, JOUR-MAS: A Multi-agent System Approach to Help Journalism Management, vol. 4, no. 4, 2015.
- P. Mozur, "A Genocide Incited on Facebook with Posts from Myanmar's Military," *The New York Times,* Oct. 15 2018. https://www.nytimes.com/2018/10/15/technology/myanmar-faceboo.k-genocide.html
- UK Parliament, House of Commons, Digital, Culture, Media and Sport Committee Disinformation and 'fake news': Interim Report, Fifth Report of Session 2017–19UK Parliament, Published on July 29, 2018.

Section 2—Equal Availability

Issue: Vastly different power structures among and within countries create risk that A/IS deployment accelerates, rather than reduces, inequality in the pursuit of a sustainable future. It is unclear how LMIC can best implement A/IS via existing resources and take full advantage of the technology's potential to achieve a sustainable future.

Background

The potential use of A/IS to create sustainable economic growth for LMIC is uniquely powerful. Yet, many of the debates surrounding A/IS take place within HIC, among highly educated and financially secure individuals. 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 inequality, classism, and general rejection of these technologies. With much of the financial and technical resources for A/IS development and deployment residing in HIC, not only are A/IS benefits more difficult to access for LMIC populations, but those A/IS applications that are deployed outside of HIC realities may not be appropriate. This is for reasons of cultural/ethnic bias, language difficulties, or simply an inability to adapt to local internet infrastructure constraints.

Furthermore, technological innovation in LMIC comes up against many potential obstacles, which could be considered when undertaking initiatives aimed at enhancing LMIC access:

- Reluctance to provide open source licensing of technological development innovations,
- Lack of the human capital and knowledge required to adapt HIC-developed technologies to resolving problems in the LMIC context, or to develop local technological solutions to these problems,
- Retention of A/IS capacity in LMIC due to globally uncompetitive salaries,
- Lack of infrastructure for deployment, and difficulties in taking technological solutions to where they are needed,
- Lack of organizational and business models for adapting technologies to the specific needs of different regions,
- Lack of active participation of the target population,
- Lack of political will to allow people to have access to technological resources,
- Existence of oligopolies that hinder new technological development,
- Lack of inclusive and high-quality education at all levels, and
- Bureaucratic policies ill-adapted to highly dynamic scenarios.

A/IS for Sustainable Development

For A/IS capacities and benefits to become equally available worldwide, training, education, and opportunities should be provided particularly for LMIC. Currently, access to products that facilitate A/IS research of timely topics is quite limited for researchers in LMIC, due to cost considerations.

If A/IS capacity and governance problems, such as relevant laws, policies, regulations, and anti-corruption safeguards, are addressed, LMIC could have the ability to use A/IS to transform their economies and leapfrog into a new era of inclusive growth. Indeed, A/IS itself can contribute to good governance when applied to the detection of corruption in state and banking institutions, one of the most serious recognized constraints to investment in LMIC. Particular attention, however, must be paid to ensure that the use of A/IS is for the common good—especially in the context of LMIC—and does not reinforce existing socioeconomic inequities through systematic discriminatory bias in both design and application, or undermine fundamental rights through, among other issues, lax data privacy laws and practice.

Recommendations

A/IS benefits should be equally available to populations in HIC and LMIC, in the interest of universal human dignity, peace, prosperity, and planet protection. Specific measures for LMIC should include:

- Deploying A/IS to detect fraud and corruption, to increase the transparency of power structures, to contribute to a favorable investment, governance, and innovation environment.
- Supporting LMIC in the development of their own A/IS strategies, and in the retention or return of their A/IS talent to prevent "brain drain".
- Encouraging global standardization/ harmonization and open source A/IS software.
- Promoting distribution of knowledge and wealth generated by the latest A/IS, including through formal public policy and financial mechanisms to advance equity worldwide.
- Developing public datasets to facilitate the access of people from LMIC to data resources to facilitate their applied research, while ensuring the protection of personal data.
- Creating A/IS international research centers in every continent, that promote culturally appropriate research, and allow the remote access of LMIC's communities to high-end technology.[16]
- Facilitating A/IS access in LMIC through online courses in local languages.
- Ensuring that, along with the use of A/IS, discussions related to identity, platforms, and blockchain are conducted, such that core enabling technologies are designed to meet the economic, social, and cultural needs of LMIC.
- Diminishing the barriers and increase LMIC access to technological products, including the formation of collaborative networks between developers in HIC and LMIC, supporting the latter in attending global A/IS conferences.[17]
- Promoting research into A/IS-based technologies, for example, mobile lightweight A/IS applications, that are readily available in LMIC.
- Facilitating A/IS research and development in LMIC through investment incentives, public-

A/IS for Sustainable Development

private partnerships, and/or joint grants, and collaboration between international organizations, government bodies, universities, and research institutes.

- Prioritizing A/IS infrastructure in international development assistance, as necessary to improve the quality and standard of living and advance progress towards the SDGs in LMIC.
- Recognizing data issues that may be particular to LMIC contexts, i.e., insufficient sample size for machine learning which sometimes results in ***de facto*** discrimination, and inadequate laws for, and the practice of, data protection.
- Supporting research on the adaptation of A/IS methods to scarce data environments and other remedies that facilitate an optimal A/IS enabling environment in LMIC.

Further Resources

- A. Akubue, "Appropriate Technology for Socioeconomic Development in Third World Countries." *The Journal of Technology Studies* 26, no. 1, pp. 33–43, 2000.
- O. Ajakaiye and M. S. Kimenyi. "Higher Education and Economic Development in Africa: Introduction and Overview." *Journal of African Economies* 20, no. 3, iii3–iii13, 2011.
- D. Allison-Hope and M. Hodge, "Artificial Intelligence: A Rights-Based Blueprint for Business," San Francisco: BSF, Aug. 28, 2018
- D. E. Bloom, D. Canning, and K. Chan. *Higher Education and Economic Development in Africa* (Vol. 102). Washington, DC: World Bank, 2006.
- N. Bloom, "Corporations in the Age of Inequality." *Harvard Business Review,* April 21, 2017.
- C. Dahlman, *Technology, Globalization, and Competitiveness: Challenges for Developing Countries. Industrialization in the 21st Century.* New York: United Nations, 2006.
- M. Fong, *Technology Leapfrogging for Developing Countries. Encyclopedia of Information Science and Technology*, 2nd ed. Hershey, PA: IGI Global, 2009 (pp. 3707– 3713).
- C. B. Frey and M. A. Osborne. "The Future of Employment: How Susceptible Are Jobs to Computerisation?" (working paper). Oxford, U.K.: Oxford University, 2013.
- B. Hazeltine and C. Bull. *Appropriate Technology: Tools, Choices, and Implications.* New York: Academic Press, 1999.
- McKinsey Global Institute. "Disruptive Technologies: Advances That Will Transform Life, Business, and the Global Economy" (report), May 2013.
- D. Rotman, "How Technology Is Destroying Jobs." *MIT Technology Review*, June 12, 2013.
- R. Sauter 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.
- World Development Report 2016: Digital Dividends. Washington, DC: World Bank. doi:10.1596/978-1-4648-0671-1.
- World Wide Web Foundation "Artificial Intelligence: The Road ahead in Low and Middle-income Countries," webfoundation.org, June 2017.

Section 3—A/IS and Employment

Issue: A/IS are changing the nature of work, disrupting employment, while 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 share 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 is 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."[18]

The risk of unemployment for LMIC is more serious than for developed countries. The industry of most LMIC is labor intensive. While labor may be cheap(er) in LMIC economies, the ripple effects of A/IS and automation will be felt much more than in the HIC economies. The 2016 World Bank Development Report stated that the share of occupations susceptible to automation and A/IS is higher in LMIC than in HIC, where such jobs have already disappeared. In addition, the qualities which made certain jobs easy to outsource to LMIC where wages are lower are those that may make them easy to automate.[19] An offsetting factor is the reality that many LMIC lack the communication, energy, and IT infrastructure required to support highly automated industries.[20] Notwithstanding this reality, the World Bank estimated the automatable share of employment, unadjusted for adoption time lag, for LMIC ranges from 85% in Ethiopia to 62% in Argentina, compared to the OECD average of 57%.[21]

In the coming decades, the automation wave calls for higher investment and the transformation of labor market capacity development programs. Innovative and fair ways of funding such an investment are required; the solutions should be designed in cooperation with the companies benefiting from the increase of profitability, thanks to automation. This should be done in a responsible way so that the innovation cycle is not broken, and yet workforce capacity does not fall behind the needs of 21st century employment. At the same time, 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 multisided platforms and predictive analytics, provided they do not entrench discrimination.[22] The case of platforms, such as LinkedIn, for instance, with its 470 million

A/IS for Sustainable Development

registered users, and online job consolidators such as indeed.com and Simply Hired, are interesting as an evolution in hiring practices, at least for those able to access the internet.

Tailored counseling and integrated retraining programs also represent promising grounds for innovation. In addition, much will have to be done to create fair and effective lifelong skill development/training, infrastructures, and mechanisms capable of empowering millions of people to viably transition jobs, sectors, and potentially locations, and to address differential geographic impacts that 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, tax incentives, and crucially in the age of A/IS, universal broadband access, especially in rural areas of both HIC and LMIC.

Recommendations

To thrive in the A/IS age, workers must be provided training in skills that improve their adaptability to rapid technological changes; programs should be available to any worker, with special attention to the low-skilled workforce. Those programs can be private, that is, sponsored by the employer, or publicly and freely offered through specific public channels and government policies, and should be available regardless of whether the worker is in between jobs or still employed. Specific measures include:

- Offering new technical programs, possibly earlier than high school, to increase the workforce capacity to close the skills gap and thrive in employment alongside A/IS.
- Creating opportunities for apprenticeships, pilot programs, and scaling up data-driven evidence-based solutions that increase employment and earnings.
- Supporting new forms of public-private partnerships involving civil society, as well as new outcome-oriented financial mechanisms, e.g., social impact bonds, that help scale up successful innovations.
- Supporting partnerships between universities, innovation labs in corporations, and governments to research and incubate startups for A/IS graduates.[23]
- Developing regulations to hold corporations responsible for employee retraining necessary due to increased automation and other technological applications having impact on the workforce.
- Facilitating private sector initiatives by public policy for co-investment in training and retraining programs through tax incentives.
- Establishing and resourcing public policies that assure the survival and well-being of workers, displaced by A/IS and automation, who cannot be retrained.
- Researching complementary areas, to lay solid foundations for the transformation outlined above.
    - Requiring more policy research on the dynamics of professional transitions in different labor market conditions.

A/IS for Sustainable Development

- Researching the fairest and most efficient public-private options for financing labor force transformation due to A/IS.
- Developing national and regional future of work strategies based on sound research and strategic foresight.

Further Resources

- V. Cerf and D. Norfors, The People-centered Economy: The New Ecosystem for Work. California: IIIJ Foundation, 2018.
- Executive Office of the President. *Artificial Intelligence, Automation, and the Economy.* December 20, 2016.
- S. Kilcarr, "Defining the American Dream for Trucking ... and the Nation, Too," *FleetOwner*, April 26, 2016.
- M. Mason, "Millions of Californians' Jobs could be Affected by Automation—a Scenario the next Governor has to Address,"*Los Angeles Times*, October 14, 2018.
- OECD, "Labor Market Programs: Expenditure and Participants," *OECD Employment and Labor Market Statistics* (database), 2016.
- M. Vivarelli, "Innovation and Employment: A Survey," Institute for the Study of Labor (IZA) Discussion Paper No. 2621, February 2007.

Issue: Analysis of the A/IS impact on employment is too focused on the number and category of jobs affected, whereas more attention should be addressed to the complexities of changing the task content of jobs.

Background

Current attention on automation and employment tends to focus on the sheer number of jobs lost or gained. It is important to focus the analysis on how employment structures will be changed by A/IS, rather than solely dwelling on the number of jobs that might be impacted. For example, rather than carrying out a task themselves, workers will need to shift to supervision of robots performing that task. Other concerns include changes in traditional employment structures, with an increase in flexible, contract-based temporary jobs, without employee protection, and a shift in task composition away from routine/repetitive and toward complex decision-making. This is in addition to the enormous need for the aforementioned retraining. Given the extent of disruption, workforce trends will need to measure time spent unemployed or underemployed, labor force participation rates, and other factors beyond simple unemployment numbers.

A/IS for Sustainable Development

The *Future of Jobs 2018* report of the World Economic Forum highlights:

"...the potential of new technologies to create as well as disrupt jobs and to improve the quality and productivity of the existing work of human employees. Our findings indicate that, by 2022, *augmentation* of existing jobs through technology may free up workers from the majority of data processing and information search tasks—and may also increasingly support them in high-value tasks such as reasoning and decision-making as augmentation becomes increasingly common over the coming years as a way to supplement and complement human labour."[24]

The report predicts the shift in skill demand between today and 2022 will be significant and that "proactive, strategic and targeted efforts will be needed to map and incentivize workforce redeployment… [and therefore]... investment decisions [on] whether to prioritize automation or augmentation and the question of whether or not to invest in workforce reskilling."[25]

Comparing Skills Demand, 2018 Versus 2022, Top Ten

TODAY, 2018	TRENDING, 2022	DECLINING, 2022
1. Analytical thinking and innovation	1. Analytical thinking and innovation	1. Manual dexterity, endurance, and precision
2. Complex problem-solving	2. Active learning and learning strategies	2. Memory, verbal, auditory, and spatial abilities
3. Critical thinking and analysis	3. Creativity, originality, and initiative	3. Management of financial and material resources
4. Active learning and learning strategies	4. Technology design and programming	4. Technology installation and maintenance
5. Creativity, originality, and initiative	5. Critical thinking and analysis	5. Reading, writing, math, and active listening
6. Attention to detail, trustworthiness	6. Complex problem-solving	6. Management of personnel
7. Emotional Intelligence	7. Leadership and social influence	7. Quality control and safety awareness
8. Reasoning, problem-solving, and ideation	8. Emotional intelligence	8. Coordination and time-management
9. Leadership and social influence	9. Reasoning, problem-solving, and ideation	9. Visual, auditory, and speech abilities
10. Coordination and time management	10. Systems analysis and evaluation	10. Technology use, monitoring, and control

Source: Future of Jobs Survey 2018, World Economic Forum, Table 4

A/IS for Sustainable Development

Recommendations

While there is evidence that robots and automation are taking jobs away in various sectors, a more balanced, granular, analytical, and objective treatment of A/IS impact on the workforce is needed to effectively inform policy making and essential workforce reskilling. Specifics to accomplish this include:

- Creating an international and independent agency able to properly disseminate objective statistics and inform the media, as well as the general public, about the impact of robotics and A/IS on jobs, tax revenue, growth,[26] and well-being.
- Analyzing and disseminating data on how current task content of jobs have changed, based on a clear assessment of the automatability of the occupational description of such jobs.
- Promoting automation with augmentation, as recommended in the *Future of Jobs Report 2018* (see chart on page 154), to maximize the benefit of A/IS to employment and meaningful work.
- Integrating more granulated dynamic mapping of the future jobs, tasks, activities, workplace-structures, associated work-habits, and skills base spurred by the A/IS revolution, in order to innovate, align, and synchronize skill development and training programs with future requirements. This workforce mapping is needed at the macro, but also crucially at the micro, levels where labor market programs are deployed.
- Considering both product and process innovation, and looking at them from a global perspective in order to understand properly the global impact of A/IS on employment.
- Proposing mechanisms for redistribution of productivity increases and developing an adaptation plan for the evolving labor market.

Further Resources

- E. Brynjolfsson and A. McAfee. The Second Age of Machine Intelligence: Work Progress and Prosperity in a Time of Brilliant Technologies. New York, NY: W. W. Norton & Company, 2014.
- P.R. Daugherty, and H.J. Wilson, Human + Machine: Reimagining Work in the Age of AI. Watertown, MA: Harvard Business Review Press, 2018.
- International Federation of Robotics. "The Impact of Robots on Productivity, Employment and Jobs," A positioning paper by the International Federation of Robotics, April 2017.
- RockEU. "Robotics Coordination Action for Europe Report on Robotics and Employment," Deliverable D3.4.1, June 30, 2016.
- World Economic Forum, Centre for the New Economy and Society, *The Future of Jobs 2018*, Geneva: WEF 2018.

Section 4—Education for the A/IS Age

Issue: Education to prepare the future workforce, in both HIC and LMIC, to design ethical A/IS applications or to have a comparative advantage in working alongside A/IS, is either lacking or unevenly available, risking inequality perpetuated across generations, within and between countries, constraining equitable growth, supporting a sustainable future, and achievement of the SDGs.

Background

Multiple international institutions, in particular educational engineering organizations,[27] have called on universities to play an active role, both locally and globally, in the resolution of the enormous problems that the world faces in securing peace, prosperity, planet protection, and universal human dignity: armed conflict, social injustice, rapid climate change, abuse of human rights, etc. Addressing global social problems is one of the central objectives of many universities, transversal to their other functions, including research in A/IS. UNESCO points out that universities' preparation of future scientists and engineers for social responsibility is presently very limited, in view of the enormous ethical and social problems associated with technology.[28] Enhancing the global dimension of engineering in undergraduate and postgraduate A/IS education is necessary, so that students can be prepared as technical professionals, aware of the opportunities and risks that A/IS present, and ready for work anywhere in the world in any sector.

Engineering studies at the university and postgraduate levels is just one dimension of the A/IS education challenge. For instance, business, law, public policy, and medical students will also need to be prepared for professions where A/IS are a partner, and to have internalized ethical principles to guide the deployment of such technologies. LMIC need financial and academic support to incorporate global A/IS professional curricula in their own universities, and all countries need to develop the pipeline by preparing elementary and secondary school students to access such professional programs. While the need for curriculum reform is recognized, the impact of A/IS on various professions and socioeconomic contexts is, at this time, both evolving and largely undocumented. Thus, the overhaul of education systems at all levels should be preceded by A/IS research.

Much of LMIC education is not globally competitive today, so there is a risk that the global advent of A/IS could negatively affect the chances of young people in LMIC finding

A/IS for Sustainable Development

productive employment, further fueling global inequality. Education systems worldwide have to be reformed and transformed to fit the new demands of the information age, in view of the changing mix of skills demanded from the workforce.[29] In 21st century education, it has been observed that children need less rote knowledge, given so much is instantly accessible on the web and more tools to network and innovate are available; less memory and more imagination should be developed; and fewer physical books and more internet access is required. Young people everywhere need to develop their capacities for creativity, human empathy, ethics, and systems thinking in order to work productively alongside robots and A/IS technologies. Science, Technology, Engineering, Art/design, and Math (STEAM) subjects need to be more extensive and more creatively taught.[30] In addition, research is needed to establish ways that a new subject, empathy, can be added to these crucial 21st century subjects in order to educate the future A/IS workforce in social skills. Instead, in rich and poor countries alike, children are continuing to be educated for an industrial age which has disappeared or never even arrived. LMIC education systems, being less entrenched in many countries, may have the potential to be more flexible than those in HIC. Perhaps A/IS can be harnessed to help educational systems to leapfrog into the 21st century, just as mobile phone technology enabled LMIC leapfrog over the phase of wired communication infrastructure.

Recommendations

Education with respect to A/IS must be targeted to three sets of students: the general public, present and future professionals in A/IS, and present and future policy makers. To prepare the future workforce to develop culturally appropriate A/IS, to work productively and ethically alongside such technologies, and to advance the UN SDGs, the curricula in HIC and LMIC universities and professional schools require innovation. Equally importantly, preuniversity education systems, starting with early childhood education, need to be reformed to prepare society for the risks and opportunities of the A/IS age, rather than the current system which prepares society for work in an industrial age that ended with the 20th century. Specific recommendations include:

- Preparing future managers, lawyers, engineers, civil servants, and entrepreneurs to work productively and ethically as global citizens alongside A/IS, through reform of undergraduate and graduate curricula as well as of preschool, primary, and secondary school curricula. This will require:
    - Fomenting interaction between universities and other actors such as companies, governments, NGOs, etc., with respect to A/IS research through definition of research priorities and joint projects, subcontracts to universities, participation in observatories, and co-creation of curricula, cooperative teaching, internships/service learning, and conferences/seminars/courses.
    - Establishing and supporting more multidisciplinary degrees that include

A/IS for Sustainable Development

A/IS, and adapting university curricula to provide a broad, integrated perspective which allows students to understand the impact of A/IS in the global, economic, environmental, and sociocultural domains and trains them as future policy makers in A/IS fields.

- Integrating the teaching of ethics and A/IS across the education spectrum, from preschool to postgraduate curricula, instead of relegating ethics to a standalone module with little direct practical application.
- Promoting service learning opportunities that allow A/IS undergraduate and graduate students to apply their knowledge to meet the needs of a community.
- Creating international exchange programs, through both private and public institutions, which expose students to different cultural contexts for A/IS applications in both HIC and LMIC.
- Creating experimental curricula to prepare people for information-based work in the 21st century, from preschool through postgraduate education.
- Taking into account transversal competencies students need to acquire to become ethical global citizens, i.e., critical thinking, empathy, sociocultural awareness, flexibility, and deontological reasoning in the planning and assessment of A/IS curricula.
- Training teachers in teaching methodologies suited to addressing challenges imposed in the age of A/IS.
- Stimulating STEAM courses in preuniversity education.
- Encouraging high-quality HIC-LMIC collaborative A/IS research in both private and public universities.

- Conducting research to support innovation in education and business for the A/IS world, which could include:
  - Researching the impact of A/IS on the governance and macro/micro strategies of companies and organizations, together with those companies, in an interdisciplinary manner which harnesses expertise of both social scientists and technology experts.
  - Researching the impact of A/IS on the business model for the development of new products and services through the collaborative efforts of management, operations, and the technical research and development function.
  - Researching how empathy can be taught and integrated into curricula, starting at the preschool level.
  - Researching how schools and education systems in low-income settings of both HIC and LMIC can leverage their less-entrenched interests to leapfrog into a 21st century-ready education system.

A/IS for Sustainable Development

- Establishing ethics observatories in universities with the purpose of fostering an informed public opinion capable of participating in policy decisions regarding the ethics and social impact of A/IS applications.
- Creating professional continuing education and employment opportunities in A/IS for current professionals, including through online and executive education courses.
- Creating educative mass media campaigns to elevate society's ongoing baseline level of understanding of A/IS systems, including what it is, if and how it can be trusted in various contexts, and what are its limitations.

Further Resources

- ABET Computing and Engineering Accreditation Criteria 2018. Available at: http://www.abet.org/accreditation/accreditation-criteria/
- ABET, 2017 ABET Impact Report, Working Together for a Sustainable Future, 2017.
- emlyon business school, Artificial Intelligence in Management (AIM) Institute http://aim.em-lyon.com
- UNESCO, *The UN Decade of Education for Sustainable Development, Shaping the Education of Tomorrow.* UNESCO 2012.

Section 5—A/IS and Humanitarian Action

Issue: A/IS are contributing to humanitarian action to save lives, alleviate suffering, and maintain human dignity both during and in the aftermath of man-made crises and natural disasters, as well as to prevent and strengthen preparedness for the occurrence of such situations. However, there are ethical concerns with both the collection and use of data during humanitarian emergencies.

Background

There have been a number of promising A/IS applications that relieve suffering in humanitarian crises, such as extending the reach of the health system by using drones to deliver blood to remote parts of Rwanda,[31] locating and removing landmines,[32] efforts to use A/IS to track movements and population survival needs following a natural disaster, and to meet the multiple management requirements of refugee camps.[33] There are also promising developments using A/IS and robotics to assist people with disabilities to recover mobility, and robots to rescue people trapped in collapsed buildings.[34] A/IS are also being used to monitor conflict zones and to enable early warning systems.[35] For example, Microsoft has partnered with the UN Human Rights Office of the High Commissioner (OHCHR) to use big data in order to track and analyze human rights violations in conflict zones.[36] Machine learning is being used for improved decision-making regarding asylum adjudication and refugee resettlement, with a view to increasing successful integration between refugees and host communities.[37] In addition, there is evidence that a recent growth in human empathy has increased well-being while diminishing psychological and physical violence,[38] inspiring some researchers to look for ways of harnessing the power of A/IS to introduce more empathy and less violence into society.

The design and ethical deployment of these technologies in crisis settings are both essential and challenging. Large volumes of both personally identifiable and demographically identifiable data are collected in fragile environments, where tracking of individuals or groups may compromise their security if data privacy cannot be assured. Consent to data use is also impractical in such environments, yet crucial for the respect of human rights.

A/IS for Sustainable Development

Recommendations

The potential for A/IS to contribute to humanitarian action to save and improve lives should be prioritized for research and development, including by organizing global research challenges, while also building in safeguards to protect the creation, collection, processing, sharing, use, and disposal of information, including data from and about individuals and populations. Specific recommendations include:

- Promoting awareness of the vulnerable condition of certain communities around the globe and the need to develop and use A/IS applications for humanitarian purposes.
- Elaborating competitions and challenges in high impact conferences and university hackathons to engage both technical and nontechnical communities in the development of A/IS for humanitarian purposes and to address social issues.
- Support civil society groups who organize themselves for the purpose of A/IS research and advocacy to develop applications to benefit humanitarian causes.[39]
- Developing and applying ethical standards for the collection, use, sharing, and disposal of data in fragile settings.
- Following privacy protection frameworks for pressing humanitarian situations that ensure the most vulnerable are protected.[40]
- Setting up clear ethical frameworks for exceptional use of A/IS technologies in life-saving humanitarian situations, compared to "normal" situations.[41]
- Stimulating the development of low-cost and open source solutions based on A/IS to address specific humanitarian problems.
- Training A/IS experts in humanitarian action and norms, and humanitarian practitioners to catalyze collaboration in designing, piloting, developing, and implementing A/IS technologies for humanitarian purposes. Forging public-private A/IS participant alliances that develop crisis scenarios in advance.
- Working on cultural and contextual acceptance of any A/IS introduced during emergencies.
- Documenting and developing quantifiable metrics for evaluating the outcomes of humanitarian digital projects, and educating the humanitarian ecosystem on the same.

A/IS for Sustainable Development

Further Resources

- E. Prestes et al., "The 2016 Humanitarian Robotics and Automation Technology Challenge [Competitions]," in *IEEE Robotics & Automation Magazine*, vol. 23, no. 3, pp. 23-24, Sept. 2016. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7565695&isnumber=7565655
- L. Marques et al., "Automation of humanitarian demining: The 2016 Humanitarian Robotics and Automation Technology Challenge," *2016 International Conference on Robotics and Automation for Humanitarian Applications (RAHA)*, Kollam, 2016, pp. 1-7. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7931893&isnumber=7931858
- CYBATHLON 2020 Preliminary Race Task Descriptions http://www.cybathlon.ethz.ch/cybathlon-2020/preliminary-race-task-descriptions.html
- CYBATHLON Scientific Publications http://www.cybathlon.ethz.ch/
- Immigration Policy Lab (IPL), "Harnessing Big Data to Improve Refugee Resettlement" https://immigrationlab.org/project/harnessing-big-data-to-improve-refugee-resettlement/
- Harvard Humanitarian Initiative, *The Signal Code*, https://signalcode.org
- J.A. Quinn, et al., "Humanitarian applications of machine learning with remote-sensing data: review and case study in refugee settlement mapping" Philosophical Transactions of the Royal Society A, 376 20170363; DOI: 10.1098/rsta.2017.0363. Aug. 6, 2018.
- Humanitarian Innovation Guide: https://higuide.elrha.org/, 2019.
- P. Meier, Digital Humanitarians: How Big Data is Changing the Face of Humanitarian Response. Florida: CRC Press, 2015.
- "Technology for human rights: UN Human Rights Office announces landmark partnership with Microsoft" https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21620&LangID=E
- M. Luengo-Oroz, "10 big data science challenges facing humanitarian organizations," UNHCR, Nov. 22, 2016. http://www.unhcr.org/innovation/10-big-data-science-challenges-facing-humanitarian-organizations/
- Optic Technologies, Press Release, Vatican Hack 2018—Results, 18 March 2018, which announced winning AI applications to benefit migrants and refugees as well as social inclusion and interfaith dialogue, http://optictechnology.org/index.php/en/news-en/151-vhack-2018winners-en

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The A/IS for Sustainable Development Committee

- **Elizabeth D. Gibbons** (Chair) – Senior Fellow and Director of the Child Protection Certificate Program, FXB Center for Health and Human Rights, Harvard T.H. Chan School of Public Health
- **Kay Firth-Butterfield** (Founding Co-Chair) – Project Head, AI and Machine Learning at the World Economic Forum. Founding Advocate of AI-Global; Senior Fellow and Distinguished Scholar, Robert S. Strauss Center for International Security and Law, University of Texas, Austin; Co-Founder, Consortium for Law and Ethics of Artificial Intelligence and Robotics, University of Texas, Austin; Partner, Cognitive Finance Group, London, U.K.
- **Raj Madhavan** (Founding Co-Chair) – Founder & CEO of Humanitarian Robotics Technologies, LLC, Maryland, U.S.A.
- **Ronald C. Arkin** – Regents' Professor & Director of the Mobile Robot Laboratory; Associate Dean for Research & Space Planning, College of Computing, Georgia Institute of Technology
- **Joanna J. Bryson** – Reader (Associate Professor), University of Bath, Intelligent Systems Research Group, Department of Computer Science
- **Renaud Champion** – Director of Emerging Intelligences, emlyon business school; Founder of Robolution Capital & CEO of PRIMNEXT
- **Chandramauli Chaudhuri** – Senior Data Scientist; Fractal Analytics
- **Rozita Dara** – Assistant Professor, Principal Investigator of Data Management and Data Governance program, School of Computer Science, University of Guelph, Canada
- **Scott L. David** – Director of Policy at University of Washington—Center for Data Management and Privacy Governance LabInformation Assurance and Cybersecurity
- **Jia He** – Executive Director of Toutiao Research (Think Tank), Bytedance Inc.
- **William Hoffman** – Associate director and head of Data-Driven Development, The World Economic Forum
- **Michael Lennon** – Senior Fellow, Center for Excellence in Public Leadership, George Washington University; Co-Founder, Govpreneur.org; Principal, CAIPP.org (Consortium for Action Intelligence and Positive Performance); Member, Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems Committee
- **Miguel Luengo-Oroz** – Chief Data Scientist, United Nations Global Pulse.

A/IS for Sustainable Development

- **Angeles Manjarrés** – Professor of the Department of Artificial Intelligence of the Spanish National Distance-Learning University
- **Nicolas Miailhe** – Co-Founder & President, The Future Society; Member, AI Expert Group at the OECD; Member, Global Council on Extended Intelligence; Senior Visiting Research Fellow, Program on Science Technology and Society at Harvard Kennedy School. Lecturer, Paris School of International Affairs (Sciences Po). Visiting Professor, IE School of Global and Public Affairs
- **Roya Pakzad** – Research Associate and Project Leader in Technology and Human Rights, Global Digital Policy Incubator (GDPi), Stanford University
- **Edson Prestes** – Professor, Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Brazil; Head, Phi Robotics Research Group, UFRGS; CNPq Fellow
- **Simon Pickin** – Professor, Dpto. de Sistemas Informáticos y Computación, Facultad de Informática, Universidad Complutense de Madrid, Spain
- **Rose Shuman** – Partner at BrightFront Group & Founder, Question Box
- **Hruy Tsegaye** – One of the founders of iCog Labs; a pioneer company in East Africa to work on Research and Development of Artificial General Intelligence, Ethiopia

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

Endnotes

[1] See, for example, the writing of T. Piketty, *Capital in the Twenty-First Century* (Cambridge: Belknap Press 2014).

[2] See preamble of the United Nations General Assembly, *Transforming our world: the 2030 Agenda for Sustainable Development* (A/RES/70/1: 21 October 2015): "*This Agenda is a plan of action for people, planet and prosperity. It also seeks to strengthen universal peace in larger freedom. We recognize that eradicating poverty in all its forms and dimensions, including extreme poverty, is the greatest global challenge and an indispensable requirement for sustainable development. All countries and all stakeholders, acting in collaborative partnership, will implement this plan. We are resolved to free the human race from the tyranny of poverty and want and to heal and secure our planet. We are determined to take the bold and transformative steps which are urgently needed to shift the world on to a sustainable and resilient path. As we embark on this collective journey, we pledge that no one will be left behind. The 17 Sustainable Development Goals and 169 targets which we are announcing today demonstrate the scale and ambition of this new universal Agenda.*"

[3] Ibid, paragraph 8.

[4] A/IS has the potential to advance positive change toward all seventeen 2030 Sustainable Development Goals, which are:

Goal 1. End poverty in all its forms everywhere

Goal 2. End hunger, achieve food security and improved nutrition and promote sustainable agriculture

Goal 3. Ensure healthy lives and promote well-being for all at all ages

Goal 4. Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all

Goal 5. Achieve gender equality and empower all women and girls

Goal 6. Ensure availability and sustainable management of water and sanitation for all

Goal 7. Ensure access to affordable, reliable, sustainable and modern energy for all

Goal 8. Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all

Goal 9. Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation

Goal 10. Reduce inequality within and among countries

Goal 11. Make cities and human settlements inclusive, safe, resilient and sustainable

Goal 12. Ensure sustainable consumption and production patterns

Goal 13. Take urgent action to combat climate change and its impacts

A/IS for Sustainable Development

Goal 14. Conserve and sustainably use the oceans, seas and marine resources for sustainable development

Goal 15. Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss

Goal 16. Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels

Goal 17. Strengthen the means of implementation and revitalize the global partnership for sustainable development

Source: United Nations General Assembly, Transforming our world: the 2030 Agenda for Sustainable Development (A/RES/70/1: 21 October 2015) p. 14

[5] United Nations Secretary General "The road to dignity by 2030: ending poverty, transforming all lives and protecting the planet" United Nations, A/69/700, 4 December 2014, pp. 25-27 http://www.un.org/ga/search/view_doc.asp?symbol=A/69/700&Lang=E

[6] Internet World Stats https://www.internetworldstats.com/stats.htm, accessed 17 May 2018.

[7] ("Internet of Things, Privacy and Security in a Connected World," FTC, https:// www.ftc.gov/system/ les/documents/reports/federal-trade-commission-staff- report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf)

[8] World Economic Forum Global Future Council on Human Rights 2016-18 "White Paper: How to Prevent Discriminatory Outcomes in Machine Learning" (WEF: March 2018).

[9] World Wide Web Foundation *Artificial Intelligence: the Road ahead in Low and Middle-income Countries* (June 2017: webfoundation.org) p.13

[10] See the Well-being chapter of *Ethically Aligned Design*, First Edition

[11] See, for example, S. Vosougi, D. Roy, and S. Aral, "The spread of true and false news online" Science 09 Mar 2018: Vol. 359, Issue 6380, pp. 1146-1151 and M. Fox, "Fake News:Lies spread faster on social media than Truth does" NBC Health News, 8 March 2018 https://www.nbcnews.com/health/health-news/fake-news-lies-spread-faster-social-media-truth-does-n854896; Cyberbullying Research Center: Summary of Cyberbullying Research 2004-2016 https://cyberbullying.org/summary-of-our-cyberbullying-research and TeenSafe "Cyberbullying Facts and Statistics" TeenSafe October 4, 2016, https://www.teensafe.com/blog/cyber-bullying-facts-and-statistics/

A. Hutchison, "Social Media Still Has a Fake News Problem ~~and~~ Digital Literacy is Largely to Blame" Social Media Today, October 5, 2018 https://www.socialmediatoday.com/news/social-media-still-has-a-fake-news-problem-and-digital-literacy-is-largel/538930/; D.D. Luxton, J.D. June, and J. M. Fairall, "Social Media and Suicide: A Public Health Perspective", Am J Public Health. 2012 May; 102(Suppl 2): S195–S200. J. Twege, T. E. Joiner, M.L. Rogers, "Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links

A/IS for Sustainable Development

to Increased New Media Screen Time" Clinical Psychological Science, November 14, 2017 https://doi.org/10.1177/2167702617723376

[12] D.D. Luxton, J.D. June, and J. M. Fairall, "Social Media and Suicide: A Public Health Perspective", Am J Public Health. 2012 May; 102(Suppl 2): S195–S200. J. Twege, T. E. Joiner, M.L. Rogers, "Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time" Clinical Psychological Science, November 14, 2017 https://doi.org/10.1177/2167702617723376

[13] T. Luong, "Thermostats, Locks and Lights: Digital Tools of Domestic Abuse." *The New York Times*, June 23, 2018, https://www.nytimes.com/2018/06/23/technology/smart-home-devices-domestic-abuse.html

[14] P. Mozur, "A Genocide incited on Facebook with posts from Myanmar's Military", *The New York Times*, October 15, 2018. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

[15] United Nations *Human Rights Council "Human rights situations that require the Council's attention Report of the independent international fact-finding mission on Myanmar*" (*A/HRC/39/64, 12 September 2018)

[16] See for example Google AI in Ghana https://www.blog.google/around-the-globe/google-africa/google-ai-ghana/

[17] See *Artificial Intelligence: the Road ahead in Low and Middle-income Countries*

[18] Executive Office of the President of the United States. *Artificial Intelligence, Automation, and the Economy.* December 20, 2016. page 21.

[19] From World Wide Web Foundation *Artificial Intelligence: The Road ahead in Low and Middle-income Countries* (June 2017: webfoundation.org) page 8.

[20] Ibid.

[21] World Bank, 2016. *World Development Report 2016: Digital Dividends.* Washington, DC: World Bank. doi:10.1596/978-1-4648-0671-1 page 129.

[22] See for example: J. Dasten, "Amazon scraps secret AI recruiting tool that showed bias against women" Reuters Business News October 9, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[23] For example, The Vector Institute, CIFAR and the Legal Innovation Group at Ryerson University. See https://vectorinstitute.ai and http://www.legalinnovationzone.ca.

[24] World Economic Forum, Centre for the New Economy and Society *the Future of Jobs 2018* (Geneva: WEF 2018) p. 3.

[25] Ibid, page 9

[26] It must be noted that the OECD is already engaged in this work as well as are some government bodies. See http://www.oecd.org/employment/future-of-work/

A/IS for Sustainable Development

[27] UNESCO, WHO, ABET, Bologna Follow-Up Group Secretariat for the European Higher Education Area

[28] UNESCO, The UN Decade of Education for Sustainable Development, Shaping the Education of Tomorrow. (UNESCO: Paris 2012).

[29] See *Future of Jobs Report 2018 Survey* table, p. 154.

[30] National Math and Science Initiative, STEM Education and Workforce, 2014 https://www.nms.org/Portals/0/Docs/STEM%20Crisis%20Page%20Stats%20and%20References.pdf

[31] https://www.bloomberg.com/news/articles/2018-08-16/this-27-year-old-launches-drones-that-deliver-blood-to-rwanda-s-hospitals

[32] https://www.theguardian.com/sustainable-business/2015/may/25/robots-rescue-lethal-rehabilitation-landmines-drones

[33] See for example, C. Fey, "Tech can improve lives in refugee camps" Cambridge Network, 10 May 2018 https://www.cambridgenetwork.co.uk/news/tech-can-improve-lives-in-refugee-camps/; https://github.com/qcri-social/AIDR/wiki/AIDR-Overview

[34] https://www.sciencemag.org/news/2017/10/searching-survivors-mexico-earthquake-snake-robots

https://www.livescience.com/48473-search-and-rescue-robot-algorithm.html

[35] http://focus.barcelonagse.eu/can-machine-learning-help-policymakers-detect-conflict/

https://www.worldbank.org/en/news/press-release/2018/09/23/united-nations-world-bank-humanitarian-organizations-launch-innovative-partnership-to-end-famine

[36] "United Nations Human Rights Office of the High Commissioner, press release, "Technology for human rights: UN Human Rights Office announces landmark partnership with Microsoft" 16 May 2017." https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21620&LangID=E

[37] For example, researchers at Stanford University are running a pilot project to develop machine learning algorithms for a better resettlement program. To train their algorithm, the Immigration Policy Lab (IPL) at Stanford University and ETH Zurich gathered data from refugee resettlement agencies in the US and Switzerland. The model is optimized based on refugees' background and skill sets to match them to a host city in which the individual has a higher chance of finding employment.

[38] See for example S. Pinker, The Better Angels of Our Nature: Why Violence has Declined (Penguin 2012) and R. Krznaric, Empathy: How it matters and how to get it. (Perigee 2015).

[39] See for example TechToronto: https://www.techtoronto.org and #AI and Big Data

[40] See for example Harvard Humanitarian Initiative Signal Code https://signalcode.org

[41] See Humanitarian Innovation Guide: https://higuide.elrha.org/

Embedding Values into Autonomous and Intelligent Systems

Society has not established universal standards or guiding principles for embedding human values and norms 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 that they are designed to adopt, learn, and follow the norms and values of the community they serve. Moreover, their actions should 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 appropriate levels of trust in A/IS in the specific contexts and roles in which A/IS function.

At the present time, the conceptual complexities surrounding what "values" are (Hitlin and Piliavin 2004[1]; Malle and Dickert 2007[2]; Rohan 2000[3]; Sommer 2016[4]) make it difficult to envision A/IS that have computational structures directly corresponding to social or cultural values such as "security," "autonomy," or "fairness". It may be a more realistic goal to embed explicit norms into such systems. Since norms are observable in human behavior, they can therefore be represented as 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 social and moral 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. For discussion of specific values that are critical for ethical considerations of A/IS, see the chapters of *Ethically Aligned Design,* "Personal Data and Individual Agency" and "Well-being".

Norms are typically expressed in terms of obligations and prohibitions, and these can be expressed computationally (Malle, Scheutz, and Austerweil 2017[5]; Vázquez-Salceda, Aldewereld and Dignum 2004[6]). They 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.

Embedding Values into Autonomous and Intelligent Systems

To address the broad objective of embedding norms and, by implication, values into A/IS, this chapter addresses three more concrete goals:

1. Identifying the norms of the specific community in which the A/IS operate,
2. Computationally implementing the norms of that community within the A/IS, and
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 the A/IS and to its users within a specific community. It is understood that there may be conflicts of values and norms when identifying, implementing, and evaluating these systems. Such conflicts are a natural part of the dynamically changing and renegotiated norm systems of any community. As a result, we advocate for an approach in which systems are designed to provide transparent signals describing the specific nature of their behavior to the individuals in the community they serve. Such signals may include explanations or offers for inspection and must be in a language or form that is meaningful to the community.

Further Resources

- S. Hitlin and J. A. Piliavin, "Values: Reviving a Dormant Concept." *Annual Review of Sociology* 30, pp.359–393, 2004.
- B. F. Malle, and S. Dickert. "Values," in *Encyclopedia of Social Psychology*, edited by R. F. Baumeister and K. D. Vohs. Thousand Oaks, CA: Sage, 2007.
- B. F. Malle, 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.
- M. J. Rohan, "A Rose by Any Name? The Values Construct." *Personality and Social Psychology Review* 4, pp. 255–277, 2000.
- U. Sommer, 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.
- J. Vázquez-Salceda, H. Aldewereld, and F. Dignum. "Implementing Norms in Multiagent Systems," in Multiagent System Technologies. MATES 2004, edited by G. Lindemann, Denzinger, I. J. Timm, and R. Unland. (Lecture Notes in Computer Science, vol. 3187.) Berlin: Springer, 2004.

Section 1—Identifying Norms for Autonomous and Intelligent Systems

We identify three issues that must be addressed in the attempt to identify norms and corresponding 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 important property of norms: their dynamically changing nature (Mack 2018[7]), 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, often in light of overarching standards, e.g., laws and international humanitarian principles.

Issue 1: Which norms should be identified?

Background

If machines engage in human communities, 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 should be identified? Laws are publicly documented and therefore easy to identify, so they can be incorporated into A/IS as long as they do not violate humanitarian or community moral principles. Social and moral norms are more difficult to ascertain, as they are expressed through behavior, language, customs, cultural symbols, and artifacts. Most important, communities ranging from families to whole nations differ to various degrees in the norms they follow. Therefore, generating a universal set of norms that applies to all A/IS in all contexts is not realistic, but neither is it advisable to completely tailor the A/IS to individual preferences. We suggest that it is feasible to identify broadly observed norms of communities in which a technology is deployed.

Furthermore, the difficulty of generating a universal set of norms is not inconsistent with the goal of seeking agreement over Universal Human Rights (see the "General Principles" chapter of *Ethically Aligned Design*). However, these universal rights are not sufficient for devising A/IS that conform to the specific norms of its community. Universal Human Rights must, however, constrain the kinds of norms that are implemented in the A/IS (cf. van de Poel 2016[8]).

Embedding norms in A/IS requires a careful understanding of the communities 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. The relevant

Embedding Values into Autonomous and Intelligent Systems

norms for self-driving vehicles, for example, may differ greatly from those for robots used in healthcare. Thus, we recommend that to develop A/IS capable of following legal, 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 and roles for which the A/IS are designed. Even when designating a narrowly defined community, e.g., a nursing home, an apartment complex, or a company, there will be variations in the norms that apply, or in their relative weighting. The norm 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 expectations and needs of a given individual, such as the arrangement of a disabled person's living accommodations. However, unique individual expectations 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.

Innovation projects and development efforts for A/IS should always rely on empirical research, involving multiple disciplines and multiple methods; to investigate and document both context- and task-specific norms, spoken and unspoken, that typically apply in a particular community. Such a set of empirically identified norms should then guide system design. This process of norm identification and implementation must be iterative and revisable. A/IS with an initial set of implemented norms may betray biases of original assessments (Misra, Zitnick, Mitchell, and Girshick 2016[9]) that can be revealed by interactions with, and feedback from, the relevant community. This leads to a process of norm updating, which is described next in Issue 2.

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 and roles that the A/IS are designed for. This norm identification process must use appropriate scientific methods and continue through the system's life cycle.

Further Resources

- Mack, Ed., "Changing social norms." *Social Research: An International Quarterly,* 85, no.1, 1–271, 2018.
- I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick, (2016). Seeing through the human reporting bias: Visual Classifiers from Noisy Human-Centric Labels. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 2930–2939. doi:10.1109/CVPR.2016.320
- I. van de Poel, "An Ethical Framework for Evaluating Experimental Technology," *Science and Engineering Ethics*, 22, no. 3,pp. 667-686, 2016.

Embedding Values into Autonomous and Intelligent Systems

Issue 2: The need for norm updating

Background

Norms are not static. They change over time, in response to social progress, political change, new legal measures, or novel opportunities (Mack 2018[10]). Norms can fade away when, for whatever reasons, fewer and fewer people adhere to them. And new norms emerge when technological innovation invites novel behaviors and novel standards, e.g., cell phone use in public.

A/IS should be equipped with a starting set of social and legal norms before they are deployed in their intended community (see Issue 1), but this will not suffice for A/IS to behave appropriately over time. A/IS or the designers of A/IS, must be adept at identifying and adding new norms to its starting set, because the initial norm identification process in the community will undoubtedly have missed some norms and because the community's norms change.

Humans rely on numerous capacities to update their knowledge of 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 or a job in a new town; and they respond to feedback from others when they exhibit uncertainty about norms or have violated a norm.

Likewise, A/IS need multiple capacities to improve their own norm knowledge and to adapt to a community's dynamically changing norms. These capacities include:

- 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 a robot to a previously unknown context or who notice the A/IS' uncertainty in a familiar context, and
- Responding to formal or informal feedback from the community when the A/IS violate 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.

We recommend that the system's norm changes be transparent. That is, the system or its designer should consult with users, designers, and community representatives when adding new norms to its norm system or adjusting the priority or content of existing norms. Allowing a system to learn new norms without public or expert review has detrimental consequences (Green and Hu 2018[11]). The form of consultation

and the specific review process will vary by machine sophistication—e.g., linguistic capacity and function/role, or a flexible social companion versus a task-defined medical robot—and best practices will have to be established. In some cases, 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 with the designer may be appropriate. In yet other cases, the A/IS may propose changes, and the relevant human community, e.g., drawn from a representative crowdsourced panel, will decide whether such changes should be implemented in the system.

Recommendation

To respond to the dynamic change of norms in society A/IS or their designers must be able to amend their norms or add new ones, while being transparent about these changes to users, designers, broader community representatives, and other stakeholders.

Further Resources

- B. Green and L. Hu. "The Myth in the Methodology: Towards a Recontextualization of Fairness in ML." Paper presented at the Debates workshop at the 35th International Conference on Machine Learning, Stockholm, Sweden 2018.
- Mack, Ed., "Changing social norms," *Social Research: An International Quarterly*, 85 (1, Special Issue), 1-271, 2018.

Issue 3: A/IS will face norm conflicts and need methods to resolve them.

Background

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 (Van den Hoven 2012[12])—must be computationally tractable by A/IS; they cannot simply stop in their 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, empirical research must 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. These more local conflict resolutions will be further constrained by some general principles, such as the "Common Good Principle" (Andre and Velasquez 1992[13]) or local and national laws. For example, a self-driving vehicle's prioritization of one factor over another in its decision-making will need to reflect the laws and norms of the population in which the A/IS are deployed, e.g., the traffic laws of a U.S. state and the United States as a whole.

Embedding Values into Autonomous and Intelligent Systems

Some priority orders can be built into a given norm network as hierarchical relations, e.g., more general prohibitions against harm to humans typically override more specific norms against lying. Other priority orders can stem from the 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), the A/IS of a racist user who demands the A/IS use derogatory language for certain social groups will have to resist such demands because community norms hierarchically override an individual user's preferences. 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 and community input, e.g., about the subtle differences between contexts, so as to refine the system's norm network (see Issue 2).

Tension may sometimes arise between a community's social and legal norms and the normative considerations of designers or manufacturers. Democratic processes may need to be developed that resolve this tension—processes that cannot be presented in detail in this chapter. Often such resolution will favor the local laws and norms, but in some cases the community may have to be persuaded to accept A/IS favoring international law or broader humanitarian principles over, say, racist or sexist local practices.

In general, we 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, the relevant community of deployment, and third-party evaluators. Just like people explain to each other why they made decisions, they will expect any A/IS to be able to explain their decisions and be sensitive to user feedback about the appropriateness of the decisions. To do so, design and development of A/IS should specifically identify the relevant groups of humans who may request explanations and evaluate the systems' behaviors. In the case of a system detecting a norm conflict, the system should consult and offer explanations to representatives from the community, e.g., randomly sampled crowdsourced members or elected officials, as well as to third-party evaluators, with the goal of discussing and resolving the norm conflict.

Recommendation

A/IS developers should identify the ways in which people resolve norm conflicts and the ways in which they expect A/IS to resolve similar norm conflicts. A system's resolution of norm conflicts must be transparent—that is, documented by the system and ready to be made available to users, the relevant community of deployment, and third-party evaluators.

Embedding Values into Autonomous and Intelligent Systems

Further Resources

- M. Velasquez, C. Andre, T. Shanks, S.J., and M. J. Meyer, "The Common Good." *Issues in Ethics*, vol. 5, no. 1, 1992.
- J. Van den Hoven, "Engineering and the Problem of Moral Overload." *Science and Engineering Ethics, vol.* 18, no. 1, pp. 143–155, 2012.
- D. Abel, 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.
- O. Bendel, 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
- S. V. Burks, 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,* vol. 58, pp. 203–217, 2012.
  - Illustrates surveys and incentivized coordination games as methods to elicit norms in a large financial services firm
- F. Cushman, V. Kumar, and P. Railton, "Moral Learning," *Cognition*, vol. 167, pp. 1–282, 2017.
- M. Flanagan, D. C. Howe, and H. Nissenbaum, "Embodying Values in Technology: Theory and Practice." *Information Technology and Moral Philosophy*, J. van den Hoven and J. Weckert, Eds., Cambridge University Press, 2008, pp. 322–53. Cambridge Core, *Cambridge University Press.* Preprint available at http://www.nyu.edu/projects/nissenbaum/papers/Nissenbaum-VID.4-25.pdf
- B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren. "Value Sensitive Design and Information Systems," in *Early Engagement and New Technologies: Opening up the Laboratory,* N. Doorn, Schuurbiers, I. van de Poel, and M. Gorman, Eds., vol. 16, pp. 55–95. Dordrecht: Springer, 2013.
  - A comprehensive introduction into Value Sensitive Design and three sample applications
- G. Mackie, 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, Sept. 2014. https://dmeforpeace.org/sites/default/files/4%2009%2030%20Whole%20What%20are%20Social%20Norms.pdf
  - A broad survey of conceptual and measurement questions regarding social norms.
- J. A. Leydens and J. C. Lucena. Engineering Justice: Transforming Engineering Education and Practice. Hoboken, NJ: John Wiley & Sons, 2018.
  - Identifies principles of engineering for social justice.

Embedding Values into Autonomous and Intelligent Systems

- B. F. Malle, "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology,* vol. 18, no. 4, pp. 243–256, 2016.
  - Discusses how a robot's norm capacity fits in the larger vision of a robot with moral competence.
- K. W. Miller, 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,* vol. 23, pp. 389–401, 2017.
  - This article raises doubts about the possibility of imbuing artificial agents with morality, or of claiming to have done so.
- Open Roboethics Initiative: www.openroboethics.org. A series of poll results on differences in human moral decision-making and changes in priority order of values for autonomous systems (e.g., on care robots), 2019.
- A. Rizzo 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,* vol. 14, pp. 335–348, 2004.
  - Describes two assessment instruments of moral reasoning (including norm maintenance) based on Kohlberg's theory of moral development.
- S. H. Schwartz, "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).
- S. H. Schwartz and K. Boehnke. "Evaluating the Structure of Human Values with Confirmatory Factor Analysis." *Journal of Research in Personality,* vol. 38, pp. 230–255, 2004.
  - Describes an older method of subjective judgments of relations among valued outcomes and a newer, formal method of analyzing these relations.
- W. Wallach 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.

Section 2—Implementing Norms in Autonomous and Intelligent Systems

Once the norms relevant to A/IS' 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.

As a general guideline, we recommend that, through the entire process of implementation of norms, designers should consider various forms and metrics of evaluation, and they should define and incorporate central criteria for assessing the A/IS' norm conformity, e.g., human-machine agreement on moral decisions, verifiability of A/IS decisions, or justified trust. In this way, implementation already prepares for the critical third phase of evaluation (discussed in Section 3).

Issue 1: Many approaches to norm implementation are currently available, and it is not yet settled which ones are most suitable.

Background

The prospect of developing A/IS 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 and intelligent systems. Following Wallach and Allen (2008)[14], we might begin to categorize these as either:

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 or unethical with respect to particular ethical requirements (Dennis,

Embedding Values into Autonomous and Intelligent Systems

Fisher, Slavkovik, Webster 2016[15]; Pereira and Saptawijaya 2016[16]; Rötzer, 2016[17]; Scheutz, Malle, and Briggs 2015[18]); or

B. Bottom-up approaches, where the system, e.g., a learning component, builds up, through experience of what is to be considered ethical and unethical in certain situations, an implicit notion of ethical behavior (Anderson and Anderson 2014[19]; Riedl and Harrison 2016[20]).

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[21].

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. Increasingly, however, A/IS will encounter situations that initially programmed norms do not clearly address, requiring algorithmic procedures to select the better of two or more novel courses of action. Recent breakthroughs in machine learning and perception enable researchers to explore bottom-up approaches in which the A/IS learn about their 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, unlike current A/IS, children can feel pain and pleasure, and empathize with others. Still, A/IS can learn to detect and take into account others' pain and pleasure, thus at least achieving some of the positive effects of empathy. As research on A/IS progresses, engineers will explore new ways to improve these capabilities.

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[22]; Winfield, Blum, and Liu 2014[23]) or "Guardians" (Etzioni 2016[24]) that monitor, restrict, and even adapt certain unacceptable behaviors proposed by the system (see 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 (Arnold, Kasenberg, and Scheutz 2017[25]).

These three architectures do not cover all possible techniques for implementing norms into A/IS. For example, some contributors to the multi-agent systems literature have integrated norms into their agent specifications (Andrighetto et al. 2013[26]), 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. 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.

Embedding Values into Autonomous and Intelligent Systems

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 and different disciplines.

Further Resources

- M. Anderson, and S. L. Anderson, "GenEth: A General Ethical Dilemma Analyzer," *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Québec City, Québec, Canada, July 27 –31, 2014, pp. 253–261, Palo Alto, CA, The AAAI Press, 2014.
- G. Andrighetto, G. Governatori, P. Noriega, and L. W. N. van der Torre, eds. Normative Multi-Agent Systems. Saarbrücken/Wadern, Germany: Dagstuhl Publishing, 2013.
- R. Arkin, "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/ Reactive Robot Architecture." *Proceedings of the 2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, Netherlands, March 12 -15, 2008, IEEE, pp. 121–128, 2008.
- T. Arnold, 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, pp. 81–88. Palo Alto, CA: The AAAI Press, 2017.
- V. Charisi, L. Dennis, M. Fisher, et al. "Towards Moral Autonomous Systems," 2017.
- A. Conn, "How Do We Align Artificial Intelligence with Human Values?" *Future of Life Institute*, Feb. 3, 2017.
- L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Formal Verification of Ethical Choices in Autonomous Systems." *Robotics and Autonomous Systems,* vol. 77, pp. 1–14, 2016.
- A. Etzioni and O. Etzioni, "Designing AI Systems That Obey Our Laws and Values." *Communications of the ACM*, vol. 59, no. 9, pp. 29–31, Sept. 2016.
- L. M. Pereira and A. Saptawijaya, Programming Machine Ethics. Cham, Switzerland: Springer International, 2016.
- M. O. Riedl and B. Harrison. "Using Stories to Teach Human Values to Artificial Agents." *AAAI Workshops 2016*. Phoenix, Arizona, February 12–13, 2016.
- F. Rötzer, ed. Programmierte Ethik: Brauchen Roboter Regeln oder Moral? Hannover, Germany: Heise Medien, 2016.
- M. Scheutz, 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.
- U. Sommer, 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.
- I. Sommerville, *Software Engineering*. Harlow, U.K.: Pearson Studium, 2001.
- W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2008.
- F. T. Winfield, 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*, M. Mistry, A. Leonardis, Witkowski, and C. Melhuish, eds. pp. 85–96. Springer, 2014.

Embedding Values into Autonomous and Intelligent Systems

Issue 2: The need for transparency from implementation to deployment

Background

When A/IS become part of social communities and behave 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 (Chaudhuri 2017[27]; Wachter, Mittelstadt, and Floridi 2017[28]), and it will allow a community to understand, predict, and modify the A/IS (see Section 1, Issue 2; for a nuanced discussion see Selbst and Barocas[29]). Moreover, as the norms embedded in A/IS are continuously updated and refined (see Section 1, Issue 2), transparency allows for appropriate trust to be developed (Grodzinsky, Miller, and Wolf 2011[30]), 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 or coder verification, and for multiple stakeholders, e.g., user, engineer, and attorney. (See IEEE P7001™, IEEE Standards Project 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 from being hacked (Kroll et al. 2016[31]). Here we briefly illustrate the broad range of transparency by reference to four ways in which systems can be transparent—traceability, verifiability, honest design, 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 Zisman2012[32]). 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 (Noble 2013[33]). (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 their use contexts (Fleischmann and Wallace 2005[34]).

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[35]). 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 (Hind et al. 2018[36]).

Embedding Values into Autonomous and Intelligent Systems

Transparency as honest design—German designer Dieter Rams coined the term "honest design" to refer to design that "does not make a product more innovative, powerful or valuable than it really is" (Vitsoe 2018[37]; see also Donelli 2015[38]; Jong 2017[39]). Honest design of A/IS is one aspect of their transparency, because it allows the user to "see through" the outward appearance and accurately infer the A/IS' actual capacities. At times, however, the physical appearance of a system does not accurately represent what the system is capable of doing—e.g., the agent displays signs of a certain human-like emotion but its internal state does not represent that human emotion. Humans are quick to make strong inferences from outward appearances of human-likeness to the mental and social capacities the A/IS might have. Demands for transparency in design therefore put a responsibility on the designer to "not attempt to manipulate the consumer with promises that cannot be kept" (Vitsoe 2018[40]).

Transparency as intelligibility—As mentioned above, humans will want to understand the A/IS' decisions and actions, especially the morally significant ones. A clear requirement for an ethical A/IS is 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[41]). Furthermore, when the system cannot explain some of its actions, technicians or designers should be available to make those actions intelligible. Along these lines, the European Union's General Data Protection Regulation (GDPR), in effect since May 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 [sic] 2016[42], for a critical discussion of this regulation.)

Recommendation

A/IS, especially those with embedded norms, must have a high level of transparency, shown as traceability in the implementation process, mathematical verifiability of their reasoning, honesty in appearance-based signals, and intelligibility of the systems' operation and decisions.

Further Resources

- d. boyd, "Transparency ≠ Accountability." *Data & Society: Points*, November 29, 2016.
- A. Chaudhuri, " Philosophical Dimensions of Information and Ethics in the Internet of Things (IoT) Technology,"The EDP Audit, Control, and Security Newsletter, vol. 56, no. 4, pp. 7-18, DOI: 10.1080/07366981.2017.1380474, 2017.
- J. Cleland-Huang, O. Gotel, and A. Zisman, eds. *Software and Systems Traceability*. London: Springer, 2012. doi:10.1007/978-1-4471-2239-5
- G. Donelli, "Good design is honest." (blog). March 13, 2015. Accessed Oct 22, 2018. https://blog.astropad.com/good-design-is-honest/
- M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM*, vol. 56, no. 9, pp. 84–93, 2013.

Embedding Values into Autonomous and Intelligent Systems

- K. R. Fleischmann and W. A. Wallace. "A Covenant with Transparency: Opening the Black Box of Models." *Communications of the ACM*, vol. 48, no. 5, pp. 93–97, 2005.
- F. S. Grodzinsky, 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*, vol. 13, pp. 17–27, 2011.
- M. Hind, et al. "Increasing Trust in AI Services through Supplier's Declarations of Conformity." *ArXiv E-Prints*, Aug. 2018. [Online] Available: https://arxiv.org/abs/1808.07261. [Accessed October 28, 2018].
- C. W. De Jong, ed., *Dieter Rams: Ten Principles for Good Design*. New York, NY: Prestel Publishing, 2017.
- J. A. Kroll, J. Huey, S. Barocas et al. "Accountable Algorithms." University of Pennsylvania Law Review 165 2017.
- S. U. Noble, "Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible." InVisible Culture 19, 2013.
- D. Selbst and S. Barocas, "The Intuitive Appeal of Explainable Machines," *87 Fordham Law Review 1085*, Available at SSRN: https://ssrn.com/abstract=3126971 or http://dx.doi.org/10.2139/ssrn.3126971, Feb. 19, 2018.
- N. Tintarev 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,* pp. 29–32, 2014.
- Vitsoe. "The Power of Good Design." *Vitsoe*, 2018. Retrieved Oct 22, 2018 from https://www.vitsoe.com/us/about/good-design.
- S.Wachter, B. Mittelstadt, and L. Floridi, "Transparent, Explainable, and Accountable AI for Robotics." Science Robotics, vol. 2, no. 6, eaan6080. doi:10.1126/scirobotics. aan6080, 2017.

Embedding Values into Autonomous and Intelligent Systems

Issue 3: Failures will occur.

Background

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 time (see Section 1, Issue 2) and interactions in the social world are particularly complex and uncertain. Thus, prevention and mitigation strategies must be adopted, and we sample four possible ones.

First, anticipating the process of evaluation 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:

- 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 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 (Oetzel and Spiekermann 2014[43]) 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 reduce 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 (Brundage et al 2018[44]; Vanderelst and Winfield 2018[45]), 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 2, 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, or greatly limiting its scope of operation, until the source of failure is identified. The fail-safe components need to be understandable, 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.

Fourth, once failures have occurred, responsible entities, e.g., corporate, government, science, and engineering, shall create a publicly accessible

Embedding Values into Autonomous and Intelligent Systems

database with undesired outcomes caused by specific A/IS systems. The database would include descriptions of the problem, background information on how the problem was detected, which context it occurred in, and how it was addressed.

In summary, we offer the following recommendation.

Recommendation

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

Further Resources

- M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfunkel, A. Dafoe, P. Scharre, T. Zeitzo, et al. " "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," CoRR abs/1802.07228 [cs.AI]. 2018. https://arxiv.org/abs/1802.07228
- M. C. Oetzel and S. Spiekermann, "A Systematic Methodology for Privacy Impact Assessments: A Design Science Approach." *European Journal of Information Systems*, vol. 23, pp. 126–150, 2014. https://link.springer.com/article/10.1057/ejis.2013.18
- D. Vanderelst and A.F. Winfield, 2018 "The Dark Side of Ethical Robots," In Proc. The First AAAI/ACM Conf. on Artificial Intelligence, Ethics and Society, New Orleans, LA, Feb. 1 -3, 2018.

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 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, among others, the quality of human-machine interactions, human approval and appreciation of the A/IS, appropriate trust in the A/IS, adaptability of the A/IS to human users, and benefits to human well-being in the presence or under the influence of the A/IS. A range of normative aspects to be considered can be found in British Standard BS 8611:2016 on Robot Ethics (British Standards Institution 2016[46]). 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 2015[47]), 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 A/IS including robots (Fisher, Dennis, and Webster 2013[48]). More general principles from system quality management may also be integrated into the evaluation process, such as the Plan-Do-Check-Act (PDCA) cycle that underlies standards like ISO 9001 (International Organization for Standardization 2015[49]).

Evaluation may be done by first parties, e.g., designers, manufacturers, and users, as well as third parties, e.g., regulators, independent testing agencies, and certification bodies. 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.

As a general guideline, we recommend that evaluation of A/IS implementations must be anticipated during a system's design, incorporated

into the implementation process, and continue throughout the system's deployment (cf. ITIL principles, BMC 2016[50]). 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. Specific issues that need to be addressed in this process are discussed next.

Further Resources

- British Standards Institution. BS8611:2016, "Robots and Robotic Devices. Guide to the Ethical Design and Application of Robots and Robotic Systems," 2016.
- BMC Software. *ITIL: The Beginner's Guide to Processes & Best Practices*. http://www.bmc.com/guides/itil-introduction.html, Dec. 6, 2016.
- M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM*, vol. 56, no. 9, pp. 84–93, 2013.
- International Organization for Standardization (2015). ISO 9001:2015, Quality management systems —Requirements. Retrieved July 12, 2018 from https://www.iso.org/standard/62085.html.
- I. Sommerville, *Software Engineering.* 10th ed. Harlow, U.K.: Pearson Studium, 2015.

Issue 1: Not all norms of a target community apply equally to human and artificial agents

Background

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 are unlikely to have identical sets of norms. People will have some unique expectations for humans than they do not for machines, e.g., norms governing the regulation of negative emotions, assuming that machines do not have such emotions. People may in some cases have unique expectations of A/IS that they do not have for humans, e.g., a robot worker, but not a human worker, is expected to work without regular breaks.

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

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. For example, a passport application AI rejected an Asian man's photo because it insisted his eyes were closed (Griffiths 2016[51]). Bias can emerge in information processing. For instance, speech recognition systems are notoriously less accurate for female speakers than for male speakers (Tatman 2016[52]). System bias can affect decisions, such as a criminal risk assessment device which overpredicts recidivism by African Americans (Angwin et al. 2016[53]). The system's bias can present itself even in its own appearance and presentation: the vast majority of humanoid robots have white "skin" color and use female voices (Riek and Howard 2014[54]).

The norm identification process detailed in Section 1 is intended to minimize individual designers' biases because the community norms are assessed empirically. The identification process also seeks to incorporate 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 assumptions. Therefore, unanticipated or undetected biases should be further reduced by including members of diverse social groups in both the planning and evaluation of A/IS and integrating community outreach into the evaluation process, e.g., DO-IT program and RRI framework. Behavioral scientists and members of the target populations will be particularly valuable when devising criterion tasks for system evaluation and assessing the success of evaluating the A/IS performance on those tasks. Such tasks would assess, for example, whether the A/IS apply norms in discriminatory ways to different races, ethnicities, genders, ages, body shapes, or to people who use wheelchairs or prosthetics, and so on.

Recommendation

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

Further Resources

- J. Angwin, J. Larson, S. Mattu, and 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.
- J. Griffiths, "New Zealand Passport Robot Thinks This Asian Man's Eyes Are Closed." CNN.com, December 9, 2016.

- L. D. Riek and D. Howard,. "A Code of Ethics for the Human-Robot Interaction Profession." *Proceedings of We Robot,* April 4, 2014.
- R. Tatman, "Google's Speech Recognition Has a Gender Bias." *Making Noise and Hearing Things*, July 12, 2016.

Issue 3: Challenges to evaluation by third parties

Background

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 the A/IS' 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[55]). 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 they matches their normative requirements, e.g., acting in accordance with relevant norms and recognizing other agents' norm violations. A "red team" may also devise scenarios that try to get the A/IS to break norms so that its vulnerabilities can be revealed.

These different evaluation techniques can be assigned different levels of "strength": strong ones demonstrate the exhaustive set of the A/IS' allowable behaviors for a range of criterion scenarios; weaker ones sample from criterion scenarios and illustrate the systems' behavior for that subsample. In the latter case, confidence in the A/IS' ability to meet normative requirements is more limited. An evaluation's concluding judgment must therefore acknowledge the strength of the verification technique used, and the expressed confidence in the evaluation—and in the A/IS themselves—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, who has the right to correct the systems, and which kind of A/IS should be subject to accountability requirements.

Embedding Values into Autonomous and Intelligent Systems

Recommendation

To maximize effective evaluation by third parties, e.g., regulators and 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 achieve accountability to the relevant communities.

Further Resources

- M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM,* vol. 56, pp. 84–93, 2013.
- K. Abney, G. A. Bekey, and P. Lin. *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: The MIT Press, 2011.
- M. Anderson and S. L. Anderson, eds. *Machine Ethics.* New York: Cambridge University Press, 2011.
- M. Boden, J. Bryson, et al. "Principles of Robotics: Regulating Robots in the Real World." *Connection Science* 29, no. 2, pp. 124–129, 2017.
- M. Coeckelbergh, "Can We Trust Robots?" *Ethics and Information Technology,* vol.14, pp. 53–60, 2012.
- L. A. Dennis, M. Fisher, N. Lincoln, A. Lisitsa, and S. M. Veres, "Practical Verification of Decision-Making in Agent-Based Autonomous Systems." *Automated Software Engineering,* vol. 23, no. 3, pp. 305–359, 2016.
- M. Fisher, 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, pp. 114–137, 2016.
- K. R. Fleischmann, *Information and Human Values*. San Rafael, CA: Morgan and Claypool, 2014.
- G. Governatori and A. Rotolo. "How Do Agents Comply with Norms? " in *Normative Multi-Agent Systems*, G. Boella, P. Noriega, G. Pigozzi, and H. Verhagen, eds., *Dagstuhl Seminar Proceedings*. Dagstuhl, Germany: Schloss Dagstuhl—Leibniz- Zentrum für Informatik, 2009.
- B. Higgins, "New York City Task Force to Consider Algorithmic Harm." *Artificial Intelligence Technology and the Law Blog*, Feb. 7, 2018. [Online]. Available: http://aitechnologylaw.com/2018/02/new-york-city-task-force-algorithmic-harm/. [Accessed Nov. 1, 2018].
- S. L. Jarvenpaa, N. Tractinsky, and L. Saarinen. "Consumer Trust in an Internet Store: A Cross-Cultural Validation" *Journal of Computer-Mediated Communication,* vol. 5, no. 2, pp. 1–37, 1999.
- E. H. Leet and W. A. Wallace. "Society's Role and the Ethics of Modeling," in *Ethics in Modeling*, W. A. Wallace, ed., Tarrytown, NY: Elsevier, 1994, pp. 242– 245.
- M. A. Mahmoud, M. S. Ahmad, M. Z. M. Yusoff, and A. Mustapha. "A Review of Norms and Normative Multiagent Systems," *The Scientific World Journal*, vol. 2014, Article ID 684587, 2014.

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Embedding Values into Autonomous Intelligent Systems Committee

- **AJung Moon** (Founding Chair) – Director of Open Roboethics Institute
- **Bertram F. Malle** (Co-Chair) – Professor, Department of Cognitive, Linguistic, and Psychological Sciences, Co-Director of the Humanity-Centered Robotics Initiative, Brown University
- **Francesca Rossi** (Co-Chair) – Full Professor, computer science at the University of Padova, Italy, currently at the IBM Research Center at Yorktown Heights, NY
- **Stefano Albrecht** – Postdoctoral Fellow in the Department of Computer Science at The University of Texas at Austin
- **Bijilash Babu** – Senior Manager, Ernst and Young, EY Global Delivery Services India LLP
- **Jan Carlo Barca** – Senior Lecturer in Software Engineering and Internet of Things (IoT), School of Info Technology, Deakin University, Australia
- **Catherine Berger** – IEEE Standards Senior Program Manager, IEEE
- **Malo Bourgon** – COO, Machine Intelligence Research Institute
- **Richard S. Bowyer** – Adjunct Senior Lecturer and Research Fellow, College of Science and Engineering, Centre for Maritime Engineering, Control and Imaging (cmeci), Flinders University, South Australia
- **Stephen Cave** – Executive Director of the Leverhulme Centre for the Future of Intelligence, University of Cambridge
- **Raja Chatila** – CNRS-Sorbonne Institute of Intelligent Systems and Robotics, Paris, France; Member of the French Commission on the Ethics of Digital Sciences and Technologies CERNA; Past President of IEEE Robotics and Automation Society
- **Mark Coeckelbergh** – Professor, Philosophy of Media and Technology, the University of Vienna
- **Louise Dennis** – Lecturer, Autonomy and Verification Laboratory, University of Liverpool
- **Laurence Devillers** – Professor of Computer Sciences, University Paris Sorbonne, LIMSI-CNRS 'Affective and social dimensions in spoken interactions'; member of the French Commission on the Ethics of Research in Digital Sciences and Technologies (CERNA)
- **Virginia Dignum** – Associate Professor, Faculty of Technology Policy and Management, TU Delft

Embedding Values into Autonomous and Intelligent Systems

- **Ebru Dogan** – Research Engineer, VEDECOM
- **Takashi Egawa** – Cloud Infrastructure Laboratory, NEC Corporation, Tokyo
- **Vanessa Evers** – Professor, Human-Machine Interaction, and Science Director, DesignLab, University of Twente
- **Michael Fisher** – Professor of Computer Science, University of Liverpool, and Director of the UK Network on the Verification and Validation of Autonomous Systems, vavas.org
- **Ken Fleischmann** – Associate Professor in the School of Information at The University of Texas at Austin
- **Edith Pulido Herrera** – Bioengineering group, Antonio Nariño University, Bogotá, Colombia
- **Ryan Integlia** – assistant professor, Electrical and Computer Engineering, Florida Polytechnic University; Co-Founder of the em[POWER] Energy Group
- **Catholijn Jonker** – Full professor of Interactive Intelligence at the Faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology. Part-time full professor at Leiden Institute of Advanced Computer Science of the Leiden University
- **Sara Jordan** – Assistant Professor of Public Administration in the Center for Public Administration & Policy at Virginia Tech
- **Jong-Wook Kim** – Professor, AI.Robotics Lab, Department of Electronic Engineering, Dong-A University, Busan, Korea
- **Sven Koenig** – Professor, Computer Science Department, University of Southern California
- **Brenda Leong** – Senior Counsel, Director of Operations, The Future of Privacy Forum
- **Alan Mackworth** – Professor of Computer Science, University of British Columbia; Former President, AAAI; Co-author of "Artificial Intelligence: Foundations of Computational Agents".
- **Pablo Noriega** – Scientist, Artificial Intelligence Research Institute of the Spanish National Research Council (IIIA-CSIC), Barcelona.
- **Rajendran Parthiban** – Professor, School of Engineering, Monash University, Bandar Sunway, Malaysia
- **Heather M. Patterson** – Senior Research Scientist, Anticipatory Computing Lab, Intel Corp.
- **Edson Prestes** – Professor, Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Brazil; Head, Phi Robotics Research Group, UFRGS; CNPq Fellow.
- **Laurel Riek** – Associate Professor, Computer Science and Engineering, University of California San Diego
- **Leanne Seeto** – Co-Founder and Strategy and Operations Precision Autonomy
- **Sarah Spiekermann** – Chair of the Institute for Information Systems & Society at Vienna University of Economics and Business; Author of the textbook "Ethical IT-Innovation", the popular book "Digitale Ethik—Ein Wertesystem für das 21. Jahrhundert" and Blogger on "The Ethical Machine"

Embedding Values into Autonomous and Intelligent Systems

- **John P. Sullins** – Professor of Philosophy, Chair of the Center for Ethics Law and Society (CELS), Sonoma State University
- **Jaan Tallinn** – Founding engineer of Skype and Kazaa; co-founder of the Future of Life Institute
- **Mike Van der Loos** – Associate Prof., Dept. of Mechanical Engineering, Director of Robotics for Rehabilitation, Exercise and Assessment in Collaborative Healthcare (RREACH) Lab, and Associate Director of CARIS Lab, University of British Columbia
- **Wendell Wallach** – Consultant, ethicist, and scholar, Yale University's Interdisciplinary Center for Bioethics
- **Nell Watson** – CFBCS, FICS, FIAP, FIKE, FRSA, FRSS, FLS Co-Founder and Chairman, EthicsNet, AI & Robotics Faculty Singularity University, Foresight Machine Ethics Fellow
- **Karolina Zawieska** – Postdoctoral Research Fellow in Ethics and Cultural Learning of Robotics at DeMontfort University, UK and Researcher at Industrial Research Institute for Automation and Measurements PIAP, Poland

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

Endnotes

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

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

[3] M. J. Rohan, "A Rose by Any Name? The Values Construct." *Personality and Social Psychology Review* 4 (2000): 255–277.

[4] A. U. Sommer, 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.

[5] B. F. Malle, 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.

[6] J. Vázquez-Salceda, H. Aldewereld, and F. Dignum. "Implementing Norms in Multiagent Systems," in Multiagent System Technologies. MATES 2004, edited by G. Lindemann, Denzinger, I. J. Timm, and R. Unland. (Lecture Notes in Computer Science, vol. 3187.) Berlin: Springer, 2004.

[7] A. Mack, (Ed.). "Changing social norms." *Social Research: An International Quarterly,* 85, no.1 (2018): 1–271.

[8] I. van de Poel, "An Ethical Framework for Evaluating Experimental Technology", *Science and Engineering Ethics*, 22, no. 3 (2016): 667-686.

[9] I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick, (2016). Seeing through the human reporting bias: Visual Classifiers from Noisy Human-Centric Labels. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 2930–2939). doi:10.1109/CVPR.2016.320

[10] A. Mack, (Ed.). (2018). Changing social norms. *Social Research: An International Quarterly*, 85(1, Special Issue), 1–271.

[11] B. Green and L. Hu. "The Myth in the Methodology: Towards a Recontextualization of Fairness in ML." Paper presented at the Debates workshop at the 35th International Conference on Machine Learning, Stockholm, Sweden 2018.

[12] J. Van den Hoven, "Engineering and the Problem of Moral Overload." *Science and Engineering Ethics* 18, no. 1 (2012): 143–155.

[13] C. Andre and M. Velasquez. "The Common Good." *Issues in Ethics* 5, no. 1 (1992).

[14] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2008.

[15] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster. "Formal Verification of Ethical Choices in Autonomous Systems." *Robotics and Autonomous Systems* 77 (2016): 1–14.

Embedding Values into Autonomous and Intelligent Systems

[16] L. M. Pereira and A. Saptawijaya. *Programming Machine Ethics*. Cham, Switzerland: Springer International, 2016.

[17] F. Rötzer, ed. Programmierte Ethik: Brauchen Roboter Regeln oder Moral? Hannover, Germany: Heise Medien, 2016.

[18] M. Scheutz, 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.

[19] M. Anderson and S. L. Anderson. "GenEth: A General Ethical Dilemma Analyzer." *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014): 253–261.

[20] M. O. Riedl 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.

[21] V. Charisi, L. Dennis, M. Fisher et al. "Towards Moral Autonomous Systems," 2017.

[22] R. Arkin, "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture." *Proceedings of the 2008 3rd ACM/IEEE International Conference on Human-Robot Interaction* (2008): 121–128.

[23] A. F. T. Winfield, 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, Witkowski, and C. Melhuish, 85–96. Springer, 2014.

[24] A. Etzioni, "Designing AI Systems That Obey Our Laws and Values." *Communications of the ACM* 59, no. 9 (2016): 29–31.

[25] T. Arnold, 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.

[26] G. Andrighetto, G. Governatori, P. Noriega, and L. W. N. van der Torre, eds. *Normative Multi-Agent Systems*. Saarbrücken/Wadern, Germany: Dagstuhl Publishing, 2013.

[27] A. Chaudhuri, (2017) Philosophical Dimensions of Information and Ethics in the Internet of Things (IoT) Technology. The EDP Audit, Control, and Security Newsletter, 56:4, 7-18, DOI: 10.1080/07366981.2017.1380474

[28] S.Wachter, B. Mittelstadt, and L. Floridi, "Transparent, Explainable, and Accountable AI for Robotics." Science Robotics 2, no. 6 (2017): eaan6080. doi:10.1126/scirobotics. aan6080

[29] A. D. Selbst and S. Barocas, The Intuitive Appeal of Explainable Machines (February 19, 2018). Fordham Law Review. Available at SSRN: https://ssrn.com/abstract=3126971 or http://dx.doi.org/10.2139/ssrn.3126971

[30] F. S. Grodzinsky, 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.

Embedding Values into Autonomous and Intelligent Systems

[31] J. A. Kroll, J. Huey, S. Barocas et al. "Accountable Algorithms." University of Pennsylvania Law Review 165 (2017).

[32] J. Cleland-Huang, O. Gotel, and A. Zisman, eds. Software and Systems Traceability. London: Springer, 2012. doi:10.1007/978- 1-4471-2239-5

[33] S. U. Noble, "Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible." InVisible Culture 19 (2013).

[34] K. R. Fleischmann and W. A. Wallace. "A Covenant with Transparency: Opening the Black Box of Models." Communications of the ACM 48, no. 5 (2005): 93–97.

[35] M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." Communications of the ACM 56, no. 9 (2013): 84–93.

[36] M. Hind, et al. "Increasing Trust in AI Services through Supplier's Declarations of Conformity." *ArXiv E-Prints*, Aug. 2018. Retrieved October 28, 2018 from https://arxiv.org/abs/1808.07261.

[37] Vitsoe. "The Power of Good Design." *Vitsoe*, 2018. Retrieved Oct 22, 2018 from https://www.vitsoe.com/us/about/good-design.

[38] G. Donelli, (2015, March 13). Good design is honest (Blogpost). Retrieved Oct 22, 2018 from https://blog.astropad.com/good-design-is-honest/

[39] C. de Jong Ed., "Ten principles for good design: Dieter Rams." New York, NY: Prestel Publishing, 2017.

[40] Ibid.

[41] N. Tintarev 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.

[42] d. boyd, "Transparency ≠ Accountability." *Data & Society: Points*, November 29, 2016.

[43] C. Oetzel and S. Spiekermann, "A Systematic Methodology for Privacy Impact Assessments: A Design Science Approach." *European Journal of Information Systems* 23, (2014): 126–150. https://link.springer.com/article/10.1057/ejis.2013.18

[44] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfunkel, A. Dafoe, P. Scharre, T. Zeitzo, et al. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. CoRR abs/1802.07228 (2018). https://arxiv.org/abs/1802.07228M.

[45] D. Vanderelst and A.F. Winfield, 2018 The Dark Side of Ethical Robots. In Proc. AAAI/ACM Conf. on Artificial Intelligence, Ethics and Society, New Orleans.

[46] British Standards Institution. BS8611:2016, "Robots and Robotic Devices. Guide to the Ethical Design and Application of Robots and Robotic Systems," 2016.

[47] I. Sommerville, *Software Engineering* (10th edition). Harlow, U.K.: Pearson Studium, 2015.

[48] M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." Communications of the ACM 56, no. 9 (2013): 84–93.

Embedding Values into Autonomous and Intelligent Systems

[49] International Organization for Standardization (2015). ISO 9001:2015, Quality management systems—Requirements. Retrieved July 12, 2018 from https://www.iso.org/standard/62085.html.

[50] BMC Software. *ITIL: The Beginner's Guide to Processes & Best Practices*. 6 Dec. 2016, http://www.bmc.com/guides/itil-introduction.html.

[51] J. Griffiths, "New Zealand Passport Robot Thinks This Asian Man's Eyes Are -Closed." CNN.com, December 9, 2016.

[52] R. Tatman, "Google's Speech Recognition Has a Gender Bias." Making Noise and Hearing Things, July 12, 2016.

[53] J. Angwin, 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.

[54] L. D. Riek and D. Howard. "A Code of Ethics for the Human-Robot Interaction Profession." Proceedings of We Robot, April 4, 2014.

[55] M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." Communications of the ACM 56 (2013): 84–93.

Policy

Introduction

Autonomous and intelligent systems (A/IS) are a part of our society. The use of these powerful technologies promotes a range of social benefits. They may spur development across economies and society through numerous applications, including in commerce, finance, employment, health care, agriculture, education, transportation, politics, privacy, public safety, national security, civil liberties, and human rights. To encourage the development of socially beneficial applications of A/IS, and to protect the public from adverse consequences of A/IS, intended or otherwise, effective policies and government regulations are needed.

Effective A/IS policies serve the public interest in several important respects. A/IS policies and regulations, at both the national level and as developed by professional organizations and governing institutions, protect and promote safety, privacy, human rights, and cybersecurity, as well as enhance the public's understanding of the potential impacts 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 hinder innovation, or regulations that do not effectively advance public interest and protect human rights.

We believe that effective A/IS policies should embody a rights-based approach[1] that addresses five issues:

1. **Ensure that A/IS support, promote, and enable internationally recognized legal norms.**
   Establish policies for A/IS using the internationally recognized legal framework for human rights standards that is directed at accounting for the impact of technology on individuals.

Policy

2. **Develop government expertise in A/IS.**
   Facilitate skill development, technical and otherwise, to further boost the ability of policy makers, regulators, and elected officials to make informed proposals and decisions about the various facets of these new technologies.

3. **Ensure governance and ethics are core components in A/IS research, development, acquisition, and use.**
   Require support for A/IS research and development (R&D) efforts with a focus on the ethical impact of A/IS. To benefit from these new technologies while also ensuring they meet societal needs and values, governments should be actively involved in supporting relevant R&D efforts.

4. **Create policies for A/IS to ensure public safety and responsible A/IS design.**
   Governments must ensure consistent and locally adaptable policies and regulations for A/IS. Effective regulation should address transparency, explainability, predictability, bias, and accountability for A/IS algorithms, as well as risk management, privacy, data protection measures, safety, and security considerations. Certification of systems involving A/IS is a key technical, societal, and industrial issue.

5. **Educate the public on the ethical and societal impacts of A/IS.**
   Industry, academia, the media, and governments must establish strategies for informing and engaging the public on benefits and challenges posed by A/IS. Communicating accurately both the positive potential of A/IS and the areas that require caution and further development is critical to effective decision-making environments.

As A/IS comprise a greater part of our daily lives, managing the associated risks and rewards becomes increasingly important. Technology leaders and policy makers have much to contribute to the debate on how to build trust, promote safety and reliability, and integrate ethical and legal considerations into the design of A/IS technologies. This chapter provides a principled foundation for these discussions.

Policy

Issue 1: Ensure that A/IS support, promote, and enable internationally recognized legal norms

Background

A/IS technologies have the potential to impact internationally recognized economic, social, cultural, and political rights through unintended outcomes and outright design decisions. Important examples of this issue have occurred with certain unmanned aircraft systems (Bowcott 2013), use of A/IS in predictive policing (Shapiro 2017), banking (Garcia 2017), judicial sentencing (Osoba and Welser 2017), and job hunting and hiring practices (Datta, Tschantz, and Datta 2014). Even service delivery of goods (Ingold and Soper 2016) can impact human rights by automating discrimination (Eubanks 2018) and inhibiting the right of assembly, freedom of expression, and access to information. To ensure A/IS are used as a force for social benefit, nations must develop policies that safeguard human rights.

A/IS regulation, development, and deployment should, therefore, be based on international human rights standards and standards of international humanitarian laws. When put into practice, both states and private actors will consider their responsibilities to protect and respect internationally recognized political, social, economic, and cultural rights. Similarly, business actors will consider their obligations to respect international human rights, as described in the United Nations Guiding Principles on Business and Human Rights (OHCHR 2011), also known as the Ruggie principles.

The Ruggie principles have been widely referenced and endorsed by corporations and have led to the adoption of several corporate social responsibility (CSR) policies in various companies. With broadened support, the Ruggie principles will strengthen the role of businesses in protecting and promoting human rights and ensuring that the most crucial human values and legal standards of human rights are respected by A/IS technologists.

Recommendations

National policies and business regulations for A/IS should be founded on a rights-based approach. The Ruggie principles provide the internationally recognized legal framework for human rights standards that accounts for the impact of technology on individuals while also addressing inequalities, discriminatory practices, and the unjust distribution of resources.

These six considerations for a rights-based approach to A/IS flow from the recommendation above:

- *Responsibility:* Identify the right holders and the duty bearers and ensure that duty bearers have an obligation to fulfill all human rights.
- *Accountability:* Oblige states, as duty bearers, to behave responsibly, to seek to represent the greater public interest, and to be open to public scrutiny of their A/IS policies.
- *Participation:* Encourage and support a high degree of participation of duty bearers, right holders, and other interested parties.

Policy

- *Nondiscrimination:* Underlie the practice of A/IS with principles of nondiscrimination, equality, and inclusiveness. Particular attention must be given to vulnerable groups, to be determined locally, such as minorities, indigenous peoples, or persons with disabilities.
- *Empowerment:* Empower right holders to claim and exercise their rights.
- *Corporate responsibility:* Ensure that companies' developments of A/IS comply with the rights-based approach. Companies must not willingly provide A/IS 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.
- O. Bowcott, "Drone Strikes by US May Violate International Law, Says UN," *The Guardian*, October 18, 2013.
- A. Shapiro, "Reform Predictive Policing," *Nature News,* vol. 541, no. 7638, pp. 458–460, Jan. 25, 2017.
- M. Garcia, "How to Keep Your AI from Turning Into a Racist Monster," *Wired*, April 21, 2017.
- O. A. Osoba, and W. Welser IV, "An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence," (Research Report 1744). Santa Monica, CA: RAND Corporation, 2017.
- A. Datta, M. C. Tschantz, and A. Datta. "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination," arXiv:1408.6491 [Cs] , 2014.
- D. Ingold, and S. Soper, "Amazon Doesn't Consider the Race of Its Customers. Should It?" *Bloomberg,* April 21, 2016.
- United Nations. Office of the High Commissioner of Human Rights. *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.
- "Mapping Regulatory Proposals for Artificial Intelligence in Europe." *Access Now*, November 2018.
- V. Eubanks, *Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press, January 2018.

Policy

Issue 2: Develop government expertise in A/IS

Background

There is a 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 experts who understand and can analyze the interactions between A/IS technologies, policy objectives, and overall societal values. Sufficient depth and breadth of technical expertise will help ensure policies and regulations successfully support innovation, adhere to national principles, and protect public safety.

Effective governance also requires an A/IS workforce that has adequate training in ethics and access to other resources on human rights standards and obligations, along with guidance on how to apply them in practice.

Recommendations

Policy makers should support the development of expertise required to create a public policy, legal, and regulatory environment that allows innovation to flourish while protecting the public and gaining public trust.[2] Example strategies include the following:

- Expertise can be furthered through technical fellowships, or rotation schemes, where technologists spend an extended time in political offices, or policy makers work with organizations[3] that operate at the intersection of technology policy, technical engineering, and advocacy. This will enhance the technical knowledge of policy makers, strengthen ties between political and technical communities, and contribute to the formulation of effective A/IS policy.
- Expertise can also be developed through cross-border sharing of best practices around A/IS legislation, consumer protection, workforce transformation, and economic displacement stemming from A/IS-based automation. This can be done through governmental cooperation, knowledge exchanges, and by building A/IS components into venues and efforts surrounding existing regulation, e.g., the General Data Protection Regulation (GDPR).
- Because A/IS involve rapidly evolving technologies, both workforce training in A/IS areas and long-term science, technology, engineering, and math (STEM) educational strategies, along with ethics courses, are needed beginning in primary school and extending into university or vocational courses. These strategies will foster A/IS expertise in the next generation of many groups, e.g., supervisors of critical systems, scientists, and policy makers.

Policy

Further Resources

- J. Holdren, and M. Smith, "Preparing for the Future of Artificial Intelligence." Washington, DC: Executive Office of the President, National Science and Technology Council, 2016.
- P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, K. Leyton-Brown, D. Parkes, W. Press, A. Saxenian, J. Shah, M. Tambe, and A. Teller. "'Artificial Intelligence and Life in 2030': One Hundred Year Study on Artificial Intelligence." (Report of the 2015-2016 Study Panel). Stanford, CA: Stanford University, 2016.
- "Japan Industrial Policy Spotlights AI, Foreign Labor." *Nikkei Asian Review,* May 20, 2016.
- Y.H. Weng, "A European Perspective on Robot Law: Interview with Mady Delvaux-Stehres." *Robohub*, July 15, 2016.

Issue 3: Ensure governance and ethics are core components in A/IS research, development, acquisition, and use.

Background

Greater national investment in ethical A/IS research and development would stimulate the economy, create high-value jobs, improve governmental services to society, and encourage international innovation and collaboration (U.S. OSTP report on the Future of AI 2016). A/IS have the potential to improve our societies through technologies such as intelligent robots and self-driving cars that will revolutionize automobile transportation and logistics systems and reduce traffic fatalities. A/IS can improve quality of life through smart cities and decision support in health care, social services, criminal justice, and the environment. To ensure such a positive effect on individuals, societies, and businesses, nations must increase A/IS R&D investments, with particular focus on the ethical development and deployment of A/IS.

International collaboration involving governments, private industry, and non-governmental organizations (NGOs) would promote the development of standards, data sharing, and norms that guide ethically aligned A/IS R&D.

Recommendations

Develop national and international standards for A/IS to enable efficient and effective public and private sector investments. Important aspects for international standards include measures of societal benefits derived from A/IS, the use of ethical considerations in A/IS investments, and risks increased or decreased by A/IS. Nations should consider their own ethical principles and develop a framework for ethics that each country could use to reflect local systems of values and laws. This will encourage actors to think both locally and globally regarding ethics. Therefore, we recommend governments to:

- Establish priorities for funding A/IS research that identify approaches and challenges for A/IS governance. This research will identify models for national and global A/IS governance and assess their benefits and adequacy to address A/IS societal needs.

Policy

- Encourage the participation of a diverse set of stakeholders in the standards development process. Standards should address A/IS issues such as fairness, security, transparency, understandability, privacy, and societal impacts of A/IS. A global framework for identification and sharing of these and other issues should be developed. Standards should incorporate independent mechanisms to properly vet, certify, audit, and assign accountability for the A/IS applications.
- Encourage and establish national and international research groups that provide incentives for A/IS research that is publicly beneficial but may not be commercially viable.

Further Resources

- E. T. Kim, "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: The National Academies Press, 1999.
- N. Chen, L. Christensen, K. Gallagher, R. Mate, and G. Rafert, "Global Economic Impacts Associated with Artificial Intelligence." Analysis Group, 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.
- S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker Project." *Proceedings of the IEEE,* vol. 102, no. 5, pp. 652–665, 2014.
- H. Markram, "The Human Brain Project," *Scientific American,* vol. 306, no. 2, pp. 50–55, June 2012.
- L. Yuan, "China Gears Up in Artificial-Intelligence Race." *Wall Street Journal*, August 24, 2016.

Issue 4: Create policies for A/IS to ensure public safety and responsible A/IS design

Background

Effective governance encourages innovation and cooperation, helps synchronize policies globally, and reduces barriers to trade. Governments must ensure consistent and appropriate policies and regulations for A/IS that address transparency, explainability, predictability, and accountability of A/IS algorithms, risk management,[4] data protection, safety, and certification of A/IS.

Appropriate regulatory responses are context-dependent and should be developed through an approach that is based on human rights[5] and has human well-being as a key goal.

Policy

Recommendations

Nations should develop and harmonize their policies and regulations for A/IS using a process that is based on informed input from a range of expert stakeholders, including academia, industry, NGOs, and government officials, that addresses questions related to the governance and safe deployment of A/IS. We recommend:

- Policy makers should consider similar work from around the world. Due to the transnational nature of A/IS, globally synchronized policies can benefit public safety, technological innovation, and access to A/IS.
- Policies should foster the development of economies able to absorb A/IS. Additional focus is needed to address the effect of A/IS on employment and income and how to ameliorate certain societal conditions. New models of public-private partnerships should be studied.
- Policies for A/IS should remain founded on a rights-based approach.
- Policy makers should be prepared to address issues that will arise when innovative and new practices enabled by A/IS are not consistent with current law. In A/IS, where there is often a different system developer, integrator, user, and ultimate customer, application of traditional legal concepts of agency, strict liability, and parental liability will require legal research and deliberation. Challenges from A/IS that must be considered include increasing complexity of and interactions between systems, and the potential for reduced predictability due to the nature of machine learning systems.

Further Resources

- P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, K. Leyton-Brown, D. Parkes, W. Press, A. Saxenian, J. Shah, M. Tambe, and A. Teller. "'Artificial Intelligence and Life in 2030': One Hundred Year Study on Artificial Intelligence." (Report of the 2015-2016 Study Panel). Stanford, CA: Stanford University, 2016.
- R. Calo, "The Case for a Federal Robotics Commission," The Brookings Institution, 2014.
- O. Groth, and Mark Nitzberg, *Solomon's Code: Humanity in a World of Thinking Machines* (chapter 8 on governance), New York: Pegasus Books, 2018.
- A. Mannes, "Institutional Options for Robot Governance," 1–40, in *We Robot 2016*, Miami, FL, April 1–2, 2016.
- G. E. Marchant, K. W. Abbott, and B. Allenby, *Innovative Governance Models for Emerging Technologies*. Cheltenham, U.K.: Edward Elgar Publishing, 2014.
- Y. H. Weng, 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, pp. 841–857, 2015.

Policy

Issue 5: Educate the public on the ethical and societal impacts of A/IS

Background

It is imperative for industry, academia, and government to communicate accurately to the public both the positive and negative potential of A/IS and the areas that require caution.[6] Strategies for informing and engaging the public on A/IS benefits and challenges are critical to creating an environment conducive to effective decision-making.

Educating users of A/IS will help influence the nature of A/IS development. Educating policy makers and regulators on the technical and legal aspects of A/IS will help enable the creation of well-defined policies that promote human rights, safety, and economic benefits. Educating corporations, researchers, and developers of A/IS on the benefits and risks to individuals and societies will enhance the creation of A/IS that better serve human well-being.[7]

Another key requirement is that A/IS are sufficiently transparent regarding implicit and explicit values and algorithmic processes. This is necessary for the public understanding of A/IS accountability, predictions, decisions, biases, and mistakes.

Recommendations

Establish an international multi-stakeholder forum, to include commercial, governmental, and other civil society groups, to determine the best practices for using and developing A/IS. Codify the deliberations into international norms and standards. Many industries—in particular, system industries (automotive, air and space, defense, energy, medical systems, manufacturing)—will be changed by the growing use of A/IS. Therefore, we recommend governments to:

- Increase funding for interdisciplinary research and communication on topics ranging from basic research on intelligence to principles of ethics, safety, privacy, fairness, liability, and trustworthiness of A/IS. Societal aspects should be addressed both at an academic level and through the engagement of business, civil society, public authorities, and policy makers.
- Empower and enable independent journalists and media outlets to report on A/IS by providing access to technical expertise.
- Conduct educational outreach to inform the public on A/IS research, development, applications, risks and rewards, along with the policies, regulations, and testing that are designed to safeguard human rights and public safety.

Policy

Develop a broad range of A/IS educational programs. Undergraduate, professional degree, advanced degree, and executive education programs should offer instruction that ensures lawyers, legislators, and A/IS workers are well informed about issues arising from A/IS, including the need for measurable standards of A/IS performance, effects, and ethics, and the need to mature the still nascent capabilities to measure these elements of A/IS.

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.
- J. Saunders, P. Hunt, and J. S. Hollywood, "Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot," *Journal of Experimental Criminology,* vol. 12, no. 347, pp. 347–371, 2016. [Online] Available: doi:10.1007/s11292-019272-0. [Accessed Nov. 10, 2018].
- B. Edelman and M. Luca, "Digital Discrimination: The Case of Airbnb.com." Harvard Business School Working Paper 14-054, Jan. 28, 2014.
- C. Garvie, A. Bedoya, and J. Frankle. "The Perpetual Line-Up: Unregulated Police Face Recognition in America." Washington, DC: Georgetown Law, Center on Privacy & Technology, 2016.
- M. Chui, and J. Manyika, "Automation, Jobs, and the Future of Work." Seattle, WA: McKinsey Global Institute, 2014.
- R. C. Arkin, "Ethics and Autonomous Systems: Perils and Promises [Point of View]." *Proceedings of the IEEE* 104, no. 10, pp. 1779–1781, Sept. 19, 2016.
- European Commission, Eurobarometer Survey on Autonomous Systems (DG Connect, June 2015), looks at Europeans' attitudes toward 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.

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Policy Committee

- **Kay Firth-Butterfield** (Founding Co-Chair) – Project Head, AI and Machine Learning at the World Economic Forum. Founding Advocate of AI-Global; Senior Fellow and Distinguished Scholar, Robert S. Strauss Center for International Security and Law, University of Texas, Austin; Co-Founder, Consortium for Law and Ethics of Artificial Intelligence and Robotics, University of Texas, Austin; Partner, Cognitive Finance Group, London, U.K.
- **Dr. Peter S. Brooks** (Co-Chair) – Institute for Defense Analyses
- **Mina Hanna** (Co-Chair) – Chair IEEE-USA Artificial Intelligence and Autonomous Systems Policy Committee, Vice Chair IEEE-USA Research and Development Policy Committee, Member of the Editorial Board of IEEE Computer Magazine
- **Chloe Autio** – Government & Policy Group, Intel Corporation
- **Stan Byers** – Frontier Markets Specialist
- **Corinne Cath-Speth** – PhD student at Oxford Internet Institute, The University of Oxford, Doctoral student at the Alan Turing Institute, Digital Consultant at ARTICLE 19
- **Michelle Finneran Dennedy** – Vice President, Chief Privacy Officer, Cisco; Author, The Privacy Engineer's Manifesto: Getting from Policy to Code to QA to Value
- **Eileen Donahoe** – Executive Director of Stanford Global Digital Policy Incubator
- **Danit Gal** – Project Assistant Professor, Keio University; Chair, IEEE Standard P7009 on the Fail-Safe Design of Autonomous and Semi-Autonomous Systems
- **Olaf J. Groth** – Professor of Strategy, Innovation, Economics & Program Director for Disruption Futures, HULT International Business School; Visiting Scholar, UC Berkeley BRIE/CITRIS; CEO, Cambrian.ai
- **Philip Hall** – (Founding Co-Chair) Co-Founder & CEO, RelmaTech; Member (and Immediate Past Chair), IEEE-USA Committee on Transportation & Aerospace Policy (CTAP); and Member, IEEE Society on Social Implications of Technology
- **John C. Havens** – Executive Director, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems; Executive Director, The Council on Extended Intelligence; Author, *Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*
- **Cyrus Hodes** – Senior Advisor, AI Office, UAE Prime Minister's Office; Co-Founded at Harvard Kennedy School the AI Initiative; Member, AI Expert Group at the OECD; Member, Global Council on Extended Intelligence.

Policy

- **Chihyung Jeon** – Assistant Professor, Graduate School of Science and Technology Policy (STP), Korea Advanced Institute of Science and Technology (KAIST)
- **Anja Kaspersen** – Former Head of International Security, World Economic Forum and head of strategic engagement and new technologies at the international committee of Red Cross (ICRC)
- **Nicolas Miailhe** – Co-Founder & President, The Future Society; Member, AI Expert Group at the OECD; Member, Global Council on Extended Intelligence; Senior Visiting Research Fellow, Program on Science Technology and Society at Harvard Kennedy School. Lecturer, Paris School of International Affairs (Sciences Po); Visiting Professor, IE School of Global and Public Affairs.
- **Simon Mueller** – Executive Director, The AI Initiative; Vice President, The Future Society
- **Carolyn Nguyen** – Director, Microsoft's Technology Policy Group, responsible for policy initiatives related to data governance and personal data
- **Mark J. Nitzberg** – Executive Director, Center for Human-Compatible Artificial Intelligence at UC Berkeley; co-author, Solomon's Code: Humanity in a World of Thinking Machines
- **Daniel Schiff** – PhD Student, Georgia Institute of Technology; Chair, Sub-Group for Autonomous and Intelligent Systems Implementation, IEEE P7010: Well-being Metric for Autonomous and Intelligent Systems
- **Evangelos Simoudis** – Co-Founder and Managing Director, Synapse Partners. Author, The Big Data Opportunity in our Driverless Future
- **Brian W. Tang** – Founder and Managing Director, Asia Capital Markets Institute (ACMI); Founding executive director, LITE Lab@HKU at Hong Kong University Faculty of Law
- **Martin Tisné** – Managing Director, Luminate
- **Sarah Villeneuve** – Policy Analyst; Member, IEEE P7010: Well-being Metric for Autonomous and Intelligent Systems
- **Adrian Weller** – Senior Research Fellow, University of Cambridge; Programme Director for AI, The Alan Turing Institute
- **Yueh-Hsuan Weng** – Assistant Professor, Frontier Research Institute for Interdisciplinary Sciences (FRIS), Tohoku University; Fellow, Transatlantic Technology Law Forum (TTLF), Stanford Law School
- **Darrell M. West** – Vice President and Director, Governance Studies | Founding Director, Center for Technology Innovation | The Douglas Dillon Chair, Brookings Institution
- **Andreas Wolkenstein** – Researcher on neurotechnologies, AI, and political philosophy at LMU Munich (Germany)

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

Endnotes

[1] This approach is rooted in internationally recognized economic, social, cultural, and political rights.

[2] 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.

[3] For example, American Civil Liberties Union, Article 19, the Center for Democracy & Technology, Canada.AI, or Privacy International. United Nations committees may also be useful in fostering knowledge exchanges.

[4] This includes consideration regarding application of the precautionary principle, as used in environmental and health policy-making, where the possibility of widespread harm is high and extensive scientific knowledge or understanding on the matter is lacking.

[5] Human rights–based approaches have been applied to development, education, and reproductive health. See the UN Practitioners' Portal on Human Rights Based Programming.

[6] "(AI100)," Stanford University., August 2016.

[7] Private sector initiatives are already emerging, such as the Partnership on AI; the AI for Good Foundation; and the Ethics and Governance of Artificial Intelligence Initiative, launched by Harvard's Berkman Klein Center for Internet & Society and the MIT Media Lab.

Law

The law affects and is affected by the development and deployment of autonomous and intelligent systems (A/IS) in contemporary life. Science, technological development, law, public policy, and ethics are not independent fields of activity that occasionally overlap. Instead, they are disciplines that are fundamentally tied to each other and collectively interact in the creation of a social order.

Accordingly, in studying A/IS and the law, we focus not only on how the law responds to the technological innovation represented by A/IS, but also on how the law guides and sets the conditions for that innovation. This interactive process is complex, and its desired outcomes can rest on particular legal and cultural traditions. While acknowledging this complexity and uncertainty, as well as the acute risk that A/IS may intentionally or unintentionally be misused or abused, we seek to identify principles that will steer this interactive process in a manner that leads to the improvement, prosperity, and well-being of everyone.

The fact that the law has a unique role to play in achieving this outcome is observed by Sheila Jasanoff, a preeminent scholar of science and technology studies:

Part of the answer is to recognize that science and technology—for all their power to create, preserve, and destroy—are not the only engines of innovation in the world. Other social institutions also innovate, and they may play an invaluable part in realigning the aims of science and technology with those of culturally disparate human societies. Foremost among these is the law.[1]

The law can play its part in ensuring that A/IS, in both design and operation, are aligned with principles of ethics and human well-being.[2]

Comprehensive coverage of all issues within our scope of study is not feasible in a single chapter of *Ethically Aligned Design (EAD)*. Accordingly, aggregate coverage will expand as issues not yet studied are selected for treatment in future versions of *EAD*.

Law

EAD, First Edition includes commentary about how the law should respond to a number of specific ethical and legal challenges raised by the development and deployment of A/IS in contemporary life. It also focuses on the impact of A/IS *on the practice of law itself*. More specifically, we study both the potential benefits and the potential risks resulting from the incorporation of A/IS into a society's legal system—specifically, in law making, civil justice, criminal justice, and law enforcement. Considering the results of those inquiries, we endeavor to identify norms for the adoption of A/IS in a legal system that will enable the realization of the benefits while mitigating the risks.[3]

In this chapter of EAD, we include the following:

Section 1: Norms for the Trustworthy Adoption of A/IS in Legal Systems.
This section addresses issues raised by the potential adoption of A/IS in legal systems for the purpose of performing, or assisting in performing, tasks traditionally carried out by humans with specialized legal training or expertise. The section begins with the question of how A/IS, if properly incorporated into a legal system, can improve the functions of that legal system and thus enhance its ability to contribute to human well-being. The section then discusses challenges to the safe and effective incorporation of A/IS into a legal system and identifies **the chief challenge as an absence of informed trust.** The remainder of the section examines how societies can fill the trust gap by enacting policies and promoting practices that advance publicly accessible standards of **effectiveness**, **competence**, **accountability**, and **transparency**.

Section 2: Legal Status of A/IS.
This section addresses issues raised by the legal status of A/IS, including the potential assignment of certain legal rights and obligations to such systems. The section provides background on the issue and outlines some of the potential advantages and disadvantages of assigning some form of legal personhood to A/IS. Based on these considerations, the section concludes that extending legal personhood to A/IS is not appropriate at this time. It then considers alternatives and outlines certain future conditions that might warrant reconsideration of the section's central recommendation.

Law

Section 1: Norms for the Trustworthy Adoption of A/IS in Legal Systems[4]

"It's a day that is here."
John G. Roberts , Chief Justice of the Supreme Court of the United States, when asked in 2017 whether he could foresee a day when intelligent machines would assist with courtroom fact-finding or judicial decision-making.[5]

A/IS hold the potential to improve the functioning of a legal system and, thereby, to contribute to human well-being. That potential will be realized, however, only if both the use of A/IS and the avoidance of their use are grounded in solid information about the capabilities and limitations of A/IS, the competencies and conditions required for their safe and effective operation (including data requirements), and the lines along which responsibility for the outcomes generated by A/IS can be assigned. Absent that information, society risks both **uninformed adoption** of A/IS and **uninformed avoidance of adoption** of A/IS, risks that are particularly acute when A/IS are applied in an integral component of the social order, such as the law.

- **Uninformed adoption** poses the risk that A/IS will be applied to inform or replace the judgments of legal actors (legislators, judges, lawyers, law enforcement officers, and jurors) without controls to ensure their safe and effective operation. They may even be used for purposes other than those for which the systems have been validated and vetted for legal use. In addition to actual harm to individuals, the result will be distrust, not only of the effectiveness of A/IS, but also of the fairness and effectiveness of the legal system itself.
- **Uninformed avoidance of adoption** poses the risk that a lack of understanding of what is required for the safe and effective operation of A/IS will result in blanket distrust of all forms and applications of A/IS, even those that are, when properly applied, safe and effective. The result will be a failure to realize the significant improvements in the legal system that A/IS can offer and a continuation of systems that are, even with the best of safeguards, still subject to human bias, inconsistency, and error.[6]

In this section, we consider how society can address these risks by developing norms for the adoption of A/IS in legal systems. The specific issues discussed follow. The first and second issues reflect the potential benefits of, and challenges to, trustworthy adoption of A/IS in the world's legal systems. The remaining issues discuss four principles,[7] which, if adhered to, will enable trustworthy adoption.[8] [9]

Law

- **Issue 1: Well-being, Legal Systems, and A/IS—**How can A/IS improve the functioning of a legal system and, thereby, enhance human well-being?
- **Issue 2: Impediments to Informed Trust—**What are the challenges to adopting A/IS in legal systems and how can those impediments be overcome?
- **Issue 3: Effectiveness—**How can the collection and disclosure of evidence of effectiveness of A/IS foster informed trust in the suitability of A/IS for adoption in legal systems?
- **Issue 4: Competence—**How can specification of the knowledge and skills required of the human operator(s) of A/IS foster informed trust in the suitability of A/IS for adoption in legal systems?
- **Issue 5: Accountability—**How can the ability to apportion responsibility for the outcome of the application of A/IS foster informed trust in the suitability of A/IS for adoption in legal systems?
- **Issue 6: Transparency—**How can sharing information that explains how A/IS reach given decisions or outcomes foster informed trust in the suitability of A/IS for adoption in legal systems?

Issue 1: Well-Being, Legal Systems, and A/IS

How can A/IS improve the functioning of a legal system and, thereby, enhance human well-being?

Background

An effective legal system contributes to human well-being. The law is an integral component of social order; the nature of a legal system informs, in fundamental ways, the nature of a society, its potential for economic growth and technological innovation, and its capacity for advancing the well-being of its members.

If the law is a constitutive element of social order, it is not surprising that it also plays a key role in setting the conditions for well-being and economic growth. In part, this flows from the fact that a well-functioning legal system is an element of good governance. Good governance and a well-functioning legal system can help society and its members flourish, as measured by indicators of both economic prosperity[10] and human well-being.[11] The attributes of good governance can be defined in several ways. Good governance can mean democracy; the observance of norms of human rights enshrined in conventions such as the Universal Declaration of Human Rights[12] and the Convention of the Rights of the Child;[13] and constitutional constraints on government power. It can also

Law

mean bureaucratic competence, law and order, property rights, and contract enforcement.

The United Nations (UN) defines the rule of law as:

> a principle of governance in which all persons, institutions and entities, public and private, including the State itself, are accountable to laws that are publicly promulgated, equally enforced and independently adjudicated. . . . It requires, as well, measures to ensure adherence to the principles of supremacy of law, equality before the law, accountability to the law, fairness in the application of the law, separation of powers, participation in decision-making, legal certainty, avoidance of arbitrariness and procedural and legal transparency.[14]

Orderly systems of legal rules and institutions generally correlate positively with economic prosperity, social stability, and human well-being, including the protection of childhood.[15] Studies from the World Bank suggest that legal reforms can lead to increased foreign investment, higher incomes, and greater wealth.[16] Wealth, in turn, can enable policies that support improved education, health, environmental protection, equal opportunity, and, in democratic societies, greater individual freedom.

Law, moreover, can contribute to prosperity not only through its functional attributes, but also through its substantive content. Patent laws, for example, if well-designed, can encourage technological innovation, leading to increases in productivity and the economic growth that follows. Poorly designed patent laws, on the other hand, may foster monopolistic markets and decrease competition, resulting in a decreased pace of technological innovation, fewer gains in productivity, and slower economic growth.[17]

While economic growth is a valuable benefit of a well-designed and well-functioning legal system, it is not the only benefit. Such a system can bring benefits to society and its members that, beyond economic prosperity, extend to mental and physical well-being. Specific benefits include the protection and advancement of an individual's dignity,[18] human rights,[19] liberty, stability, security, equality of treatment under the law, and ability to provide for the future.[20]

In fact, recent thinking on the relationship between law and economic development has come to hold that a well-functioning legal system is not simply a *means* to development but *is* development, insofar as such a system is a constitutive element of a social order that protects and advances human dignity, rights, and well-being. As this position has been characterized by David Kennedy:

> … the focal point for development policy was increasingly provided less by economics than from ideas about the nature of the good state themselves provided by literatures of political science, political economy, ethics, social theory, and law. In particular, "human rights" and the "rule of law"[21] became substantive definitions of development. One should promote human rights not to *facilitate* development—but *as* development. The rule of law was not a development *tool*—it was itself a development

Law

objective. Increasingly, law—understood as a combination of human rights, courts, property rights, formalization of entitlements, prosecution of corruption, and public order—came to define development.[22]

While this shift from considering law as a means to an end to considering law as an end in itself has been criticized on the grounds that it takes the focus off the difficult political choices that are inherent in any development policy,[23] it remains true that a well-functioning legal system is essential to the realization of a social order that protects and advances human dignity, rights, and well-being.

A/IS can contribute to the proper functioning of a legal system. A properly functioning legal system, one that is conducive to both economic prosperity and human well-being, will have a number of attributes. It should be:

- **Speedy:** enable quick resolution of civil and criminal cases;
- **Fair:** produce results that are just and proportionate to circumstance;[24]
- **Free from undesirable bias:** operate without prejudice;
- **Consistent:** arrive at outcomes in a principled, consistent, and nonarbitrary manner;
- **Transparent:** be open to appropriate public examination and oversight;[25]
- **Accessible:** be equally open to all citizens and residents in resolving disputes;
- **Effective:** achieve the ends intended by its laws and rules without negative collateral consequences;[26]
- **Accurate:** achieve accurate results, minimizing both false positives (persons unjustly or incorrectly targeted, investigated, or sentenced for crimes) and false negatives (persons incorrectly *not* targeted, investigated, or sentenced for crimes);
- **Adaptable:** have the flexibility to adapt to changes in societal circumstances.

A/IS have the potential to alter the overall functioning of a legal system. A/IS, applied responsibly and appropriately, could improve the legislative process, enhance access to justice, accelerate judicial decision-making, provide transparent and readily accessible information on why and how decisions were reached, reduce bias, support uniformity in judicial outcomes, help society identify (and potentially correct) judicial errors, and improve public confidence in the legal system. By way of example:

- A/IS can make legislation and regulation more **effective** and **adaptable**. For lawmaking, A/IS could help legislators analyze data to craft more finely tuned, responsive, evidence-based laws and regulations. This could, potentially, offer self-correcting suggestions to legislators (and to the general public) to help inform dialogue on how to meet defined public policy objectives.
- A/IS can make the practice of law more effective and efficient. For example, A/IS can enhance the **speed**, **accuracy**, and **accessibility** of the process of fact-finding in legal proceedings. When used appropriately in legal fact-finding, particularly in jurisdictions that allow extensive discovery or disclosure, A/IS already make litigation and investigations more accessible by analyzing vast data

Law

collections faster, more efficiently, and potentially more effectively[27] than document analysis conducted solely by human attorneys. By making fact-finding in an era of big data progressively easier, faster, and cheaper, A/IS may facilitate access to justice for parties that otherwise may find using the legal system to resolve disputes cost-prohibitive. A/IS can also help ensure that justice is rendered based on better accounting of the facts, thus serving the central purpose of any legal system.

- In both civil and criminal proceedings, A/IS can be used to improve the **accuracy, fairness,** and **consistency** of decisions rendered during proceedings. A/IS could serve as an auditing function for both the civil and criminal justice systems, helping to identify and correct judicial and law enforcement errors.[28]
- A/IS can increase the **speed**, **accuracy**, **fairness**, **freedom from bias**, and general **effectiveness** with which law enforcement resources are deployed to combat crime. A/IS could be used to reduce or prevent crime, respond more quickly to crimes in progress, and improve collaboration among different law enforcement agencies.[29]
- A/IS can help ensure that determinations about the arrest, detention, and incarceration of individuals suspected of, or convicted of, violations of the law are **fair**, **free from bias**, **consistent**, and **accurate**. Automated risk assessment tools have the potential to address issues of systemic racial bias in sentencing, parole, and bail determination while also safely reducing incarceration and recidivism rates by identifying individuals who are less likely to commit crimes if released.
- A/IS can help to ensure that the tools, procedures, and resources of the legal system are more **transparent** and **accessible** to citizens. For the ordinary citizen, A/IS can democratize access to legal expertise, especially in smaller matters, where they may provide effective, prompt, and low-cost initial guidance to an aggrieved party; for example, in landlord-tenant, product purchase, employment, or other contractual contexts where the individual often tends to find access to legal information and legal advice prohibitive, or where asymmetry of resources between the parties renders recourse to the legal system inequitable.[30]

A/IS have the potential to improve how a legal system functions in fundamental ways. As is the case with all powerful tools, there are some risks. **A/IS should not be adopted in a legal system without due care and scrutiny;** they should be adopted after a society's careful reflection and proper examination of evidence that their deployment and operation can be trusted to advance human dignity, rights, and well-being (see Issues 2–6).

Recommendations[31]

1. Policymakers should, in the interest of improving the function of their legal systems and bringing about improvements to human well-being, explore, through a broad consultative dialogue with all stakeholders, how A/IS can be adopted for use in their legal systems. They should do

so, however, only in accordance with norms for adoption that mitigate the risks attendant on such adoption (see Issues 2–6 in this section).

2. Governments, non-governmental organizations, and professional associations should support educational initiatives designed to create greater awareness among all stakeholders of the potential benefits and risks of adopting A/IS in the legal system, and of the ways of mitigating such risks. A particular focus of these initiatives should be the ordinary citizen who interacts with the legal system as a victim or criminal defendant.

Further Resources

- A. Brunetti, G. Kisunko, and B. Weder, "Credibility of Rules and Economic Growth: Evidence from a Worldwide Survey of the Private Sector," The World Bank Economic Review, vol. 12, no. 3, pp. 353-384, Sep. 1998.
- S. Jasanoff, "Governing Innovation: The Social Contract and the Democratic Imagination," Seminar, vol. 597, pp. 16-25, May 2009.
- D. Kennedy, "The 'Rule of Law,' Political Choices and Development Common Sense," in *The New Law and Economic Development: A Critical Appraisal*, D. M. Trubek and A. Santos, eds., Cambridge: Cambridge University Press, 2006, pp. 95-173.
- "Artificial Intelligence," National Institute of Standards and Technology.
- K. Schwab, "The Global Competitiveness Report: 2018," The World Economic Forum, 2018.
- A. Sen, Development as Freedom. New York, NY: Alfred A. Knopf, 1999.
- United Nations General Assembly, Universal Declaration of Human Rights, Dec. 10, 1948.
- UNICEF, Convention on the Rights of the Child, Nov. 4, 2014.
- United Nations Office of the High Commissioner: Human Rights, The Vienna Declaration and Programme of Action, June 25, 1993.
- World Bank, World Development Report 2017: Governance and the Law, Jan. 2017.
- World Justice Project, Rule of Law Index, June 2018.

Law

Issue 2: Impediments to Informed Trust

What are the challenges to adopting A/IS in legal systems and how can those impediments be overcome?

Background

Although the benefits to be gained by adopting A/IS in legal systems are potentially numerous (see the discussion of Issue 1), there are also significant risks that must be addressed in order for the A/IS to be adopted in a manner that will realize those benefits. The risks sometimes mirror expected benefits:

- the potential for opaque decision-making;
- the intentional or unintentional biases and abuses of power;
- the emergence of nontraditional bad actors;
- the perpetuation of inequality;
- the depletion of public trust in a legal system;
- the lack of human capital active in judicial systems to manage and operate A/IS;
- the sacrifice of the spirit of the law in order to achieve the expediency that the letter of the law allows;
- the unanticipated consequences of the surrender of human agency to nonethical agents;
- the loss of privacy and dignity;
- and the erosion of democratic institutions.[32]

By way of example:

- Currently, A/IS used in justice systems are not subject to uniform rules and norms and are often adopted piecemeal at the local or regional level, thereby creating a highly variable landscape of tools and adoption practices. Critics argue that, far from improving fact-finding in civil and criminal matters or eliminating bias in law enforcement, these tools have unproven accuracy, are error-prone, and may serve to entrench existing social inequalities. These tools' potential must be weighed against their pitfalls. These include unclear efficacy; incompetent operation; and potential impairment of a legal system's ability to adhere to principles of socioeconomic, racial, or religious equality, government transparency, and individual due process, to render justice in an informed, consistent, and fair manner.
- In the case of *State v. Loomis*, an important but not widely known case, the Wisconsin Supreme Court held that a trial court's use of an algorithmic risk assessment tool in sentencing did not violate the defendant's due process rights, despite the fact that the methodology used to obtain the automated assessment was not disclosed to either the court or the defendant.[33] A man received a lengthy sentence based in part on what an opaque algorithm thought of him. While the court considered many factors, and sought to balance competing societal values, this

Law

is just one case in a growing set of cases illustrating how criminal justice systems are being impacted by proprietary claims of trade secrets, opaque operation of A/IS, a lack of evidence of the effectiveness of A/IS, and a lack of norms for the adoption of A/IS in the extended legal system.

- More generally, humans tend to be subject to the cognitive bias known as "anchoring", which can be described as the excessive reliance on an initial piece of information. This may lead to the progressive, unwitting, and detrimental reliance of judges and legal practitioners on assessments produced by A/IS. This risk is compounded by the fact that A/IS are (and shall remain in the foreseeable future) nonethical agents, incapable of empathy, and thus at risk of being unable to produce decisions aligned with not just the letter of the law, but also the spirit of the law and reasonable regard for the circumstances of each defendant.
- The required technical and scientific knowledge to procure, deploy, and effectively operate A/IS, as well as that required to measure the ability of A/IS to achieve a given purpose without adverse collateral consequences, represent significant hurdles to the beneficial long-term adoption of A/IS in a legal system. This is especially the case when—as is the case presently—actors in the civil and criminal justice systems and in law enforcement may lack the requisite specialized technological or scientific expertise.[34]

Such risks must be addressed in order to ensure sustainable management and public oversight of what will foreseeably become an increasingly automated justice system.[35] The view expressed by the Organisation for Economic Co-operation and Development (OECD) in the domain of digital security that "robust strategies to [manage risk] are essential to establish the trust needed for economic and social activities to fully benefit from digital innovation"[36] applies equally to the adoption of A/IS in the world's legal systems.

Informed trust. If we are to realize the benefits of A/IS, we must trust that they are safe and effective. People board airplanes, take medicine, and allow their children on amusement park rides because they trust that the tools, methods, and people powering those technologies meet certain safety and effectiveness standards that reduce the risks to an acceptable level given the objectives and benefits to be achieved. This need for trust is especially important in the case of A/IS used in a legal system. The "black box" nature and lack of trust in A/IS deployed in the service of a legal system could quickly translate into a lack of trust in the legal system itself. This, in turn, may lead to an undermining of the social order. Therefore, if we are to improve the functioning of our legal systems through the adoption of A/IS, **we must enact policies and promote practices that allow those technologies to be adopted on the basis of informed trust**. Informed trust rests on a reasoned evaluation of clear and accurate information about the effectiveness of A/IS and the competence of their operators.[37]

Law

To formulate policies and standards of practice intended to foster informed trust, it is helpful, first, to identify principles applicable over the entire supply chain for the delivery of A/IS-enabled decisions and guidance, including design, development, procurement, deployment, operation, and validation of effectiveness, that, if adhered to, will foster trust. Once those general principles have been identified, specific policies and standards of practice can be formulated that encourage adherence to the principles in every aspect of a legal system, including lawmaking, civil and criminal justice, and law enforcement. Such principles, if they are to serve their intended purpose of informing effective policies and practices, must meet certain design criteria. Specifically, **the principles should be (a) individually necessary and collectively sufficient, (b) globally applicable but culturally flexible, and (c) capable of being operationalized in applicable functions of the legal system**. A set of principles that meets these criteria will provide an effective framework for the development of policies and practices that foster trust, while leaving considerable flexibility in the specific policies and standards of practice that a society chooses to implement in furthering adherence to the principles.

A set of four principles that we believe meets the design criteria just described are the following:

- **Effectiveness**: Adoption of A/IS in a legal system should be based on sound empirical evidence that they are fit for their intended purpose.
- **Competence**: A/IS should be adopted in a legal system only if their creators specify the skills and knowledge required for their effective operation and if their operators adhere to those competency requirements.
- **Accountability**: A/IS should be adopted in a legal system only if all those engaged in their design, development, procurement, deployment, operation, and validation of effectiveness maintain clear and transparent lines of responsibility for their outcomes and are open to inquiries as may be appropriate.
- **Transparency**: A/IS should be adopted in a legal system only if the stakeholders in the results of A/IS have access to pertinent and appropriate information about their design, development, procurement, deployment, operation, and validation of effectiveness.

In the remainder of Section 1, we elaborate on each of these principles. Before turning to a specific discussion of each, we add two further considerations that should be kept in mind when applying them collectively.

Differences in emphasis. While all four of the aforementioned principles will contribute to the fostering of trust, each principle will *not* contribute equally in every circumstance. For example, in many applications of A/IS, a well-established measure of effectiveness, obtained by proven and accepted methods, may go a considerable way to creating conditions for trust in the given application. In such a case, the other principles may add to trust, but they may not be necessary to establish trust. Or, to take another example, in some applications the role of the human operator may be minimal, while in other applications there will be extensive scope for

Law

human agency where competence has a greater role to play. In finding the right emphasis and balance among the four principles, policymakers and practitioners will have to consider the specific circumstances of A/IS.

Flexibility in implementation. It should be noted that we have addressed the four principles above at a rather high level and have not offered specific prescriptions of how adherence to the principles should be implemented. This is by design. Although adherence to all four principles is important, it is also important that, at the operational level, flexibility be allowed for the selection and implementation of policies and practices that (a) are in harmony with a given society's traditions, norms, and values; (b) conform with the laws and regulations operative in a given jurisdiction; and (c) are consistent with the ethical obligations of legal practitioners.

Recommendations

1. Governments should set procurement and contracting requirements that encourage parties seeking to use A/IS in the conduct of business with or for the government, particularly with or for the court system and law enforcement agencies, to adhere to the principles of effectiveness, competence, accountability, and transparency as described in this chapter. This can be achieved through legislation or administrative regulation. All government efforts in this regard should be transparent and open to public scrutiny.
2. Professionals engaged in the practice, interpretation, and enforcement of the law (such as lawyers, judges, and law enforcement officers), when engaging with or relying on providers of A/IS technology or services, should require, at a minimum, that those providers adhere to, and be able to demonstrate adherence to, the principles of effectiveness, competence, accountability, and transparency as described in this chapter. Likewise, those professionals, when operating A/IS themselves, should adhere to, and be able to demonstrate adherence to, the principles of effectiveness, competence, accountability, and transparency. Demonstrations of adherence to the requirements should be publicly accessible.
3. Regulators should permit insurers to issue professional liability and other insurance policies that consider whether the insured (either a provider or operator of A/IS in a legal system) adheres to the principles of effectiveness, competence, accountability, and transparency (as they are articulated in this chapter).

Further Resources

- "Criminal Law—Sentencing Guidelines—Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing—State v. Loomis, 881 N.W.2d 749 (Wis. 2016)," Harvard Law Review, vol. 130, no. 5, pp. 1530-1537, 2017.
- K. Freeman, "Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis," North Carolina Journal of Law and Technology, vol. 18, no. 5, pp. 75-76, 2016.

- "Managing Digital Security and Privacy Risk: Background Report for Ministerial Panel 3.2," Organisation for Economic Co-operation and Development (OECD) Directorate for Science, Technology, and Innovation: Committee on Digital Economy Policy, June 1, 2016.
- *State v Loomis*, 881 N.W.2d 749 (Wis. 2016), *cert. denied* (2017).
- "Global Governance of AI Roundtable: Summary Report 2018," World Government Summit, 2018.

Issue 3: Effectiveness

How can the collection and disclosure of evidence of effectiveness of A/IS foster informed trust in the suitability for adoption in legal systems?

Background

An essential component of trust in a technology is trust that it works and meets the purpose for which it is intended. We now turn to a discussion of the role that evidence of effectiveness, chiefly in the form of the results of a measurement exercise, can play in fostering informed trust in A/IS as applied in legal systems.[38] We begin with a general characterization of what we mean by *evidence of effectiveness*: what we are measuring, how we are measuring, what form our results take, and who the intended consumers of the evidence are. We then identify the specific features of the practice of measuring effectiveness that will enable it to contribute to informed trust in A/IS as applied in a legal system.

What constitutes evidence of effectiveness?

What we are measuring. In gathering evidence of effectiveness, we are seeking to gather empirical data that will tell us whether a given technology or its application will serve as an effective solution to the problem it is intended to address. Serving as an effective solution means more than meeting narrow specifications or requirements; it means that **the A/IS are capable of addressing their target problems in the real world**, which, in the case of A/IS applied in a legal system, are problems in the making, administration, adjudication, or enforcement of the law. It also means remaining practically feasible once collateral concerns and potential unintended consequences are taken into account.[39] To take a non-A/IS example, under the definition of effectiveness we are considering, for an herbicide to be considered effective, it must be shown not only to kill the target weeds, but also to do so without causing harm to nontarget plants, to the person applying the agent, and to the environment in general.

Under the definition above, assessing the effectiveness of A/IS in accomplishing the target task (narrowly defined) is not sufficient; it may also be necessary to assess the extent to which the A/IS are aligned with applicable

Law

laws, regulations, and standards,[40] and whether (and to what extent) they impinge on values such as privacy, fairness, or freedom from bias.[41] Whether such collateral concerns are salient will depend on the nature of the A/IS and on the particular circumstances in which they are to be applied.[42] However, it is only from such a complete view of the impact of A/IS that a balanced judgment can be made of the appropriateness of their adoption.[43]

Although the scope of an evaluation of effectiveness is broader than a narrowly focused verification that a specific requirement is met, it has its limits. There are measures of aspects of A/IS that one might find useful but that are outside the scope of effectiveness. For example, given frequently expressed concerns that A/IS will one day cross the limits of their intended purpose and overwhelm their creators and users, one might seek to define and obtain general measures of the autonomy of a system or of a system's capacity for artificial general intelligence (AGI). Although such measures could be useful—assuming they could be defined—they are beyond the scope of evaluations of effectiveness. Effectiveness is always tied to a target purpose, even if it includes consideration of the collateral effects of the manner of meeting that purpose.

What we are measuring is therefore a general "fitness for purpose".

How we measure. Evidence of effectiveness is typically gathered in one of two types of exercises:[44]

- **A single-system validation exercise** measures and reports on the effectiveness of a single system on a given task. In such an exercise, the system to be validated will typically have already carried out the target task on a given data set. The purpose of the validation is to provide empirical evidence of how successful the system has been in carrying out the task on that data set. Measurements are obtained by independent sampling and review of the data to which the system was applied. Once obtained, those metrics serve to corroborate or refute the hypothesis that the system operated as intended in the instance under consideration. An example of validation as applied to legal fact-finding would be a test of the effectiveness of A/IS that had been used to retrieve material relevant (as defined by the humans deploying the system) to a given legal inquiry from a collection of emails.
- **A multi-system (or benchmarking) evaluation** involves conducting a comparative study of the effectiveness of several systems designed to meet the same objective. Typically, in such a study, a test data set is identified, a task to be performed is defined (ideally, a task that models the real-world objectives and conditions for which the systems under evaluation have been designed[45]), the systems to be evaluated are used to carry out the task, and the success of each system in carrying out the task is measured and reported. An example of this sort of evaluation applied to a specific

real-world challenge in the justice system is the series of evaluations of the effectiveness of information retrieval systems in civil discovery, including A/IS, conducted as part of the US National Institute of Standards and Technology (NIST) Text REtrieval Conference (TREC) Legal Track initiative.[46]

The measurements obtained by both types of evaluation exercises are valuable. The results of a single-system validation exercise are typically more specific, answering the question of whether a system *was* effective in a specific instance. The results of a multi-system evaluation are typically more generic, answering the question of whether a system *can* be effective in real-world circumstances. Both questions are important, hence both types of evaluations are valuable.[47]

The form of results. The results of an evaluation typically take the form of a number—a quantitative gauge of effectiveness. This can be, for example, the decreased likelihood of developing a given medical condition; safety ratings for automobiles; recall measures for retrieving responsive documents; and so on. Certainly, qualitative considerations are not (and should not) be ignored; they often provide context crucial to interpreting the quantitative results.[48] Nevertheless, at the heart of the results of an evaluation exercise is a number, a metric that serves as a telling indicator of effectiveness.[49]

In some cases, the research community engaged in developing any new system will have reached consensus on salient effectiveness metrics. In other cases, the research community may not have reached a consensus, requiring further study. In the case of A/IS, given both their accelerating development and the fact that they are often applied to tasks for which the effectiveness of their human counterparts is seldom precisely gauged, we are often still at the stage of defining metrics. An example of an application of A/IS for which there is a general consensus around measures of effectiveness is legal electronic discovery,[50] where there is a working consensus around the use of the evaluation metrics referred to as "recall" and "precision".[51] Conversely, in the case of A/IS applied in support of sentencing decisions, a consensus on the operative effectiveness metrics does not yet exist.[52]

The consumers of the results. In defining metrics, it is important to keep in mind the consumers of the results of an evaluation of effectiveness. Broadly speaking, it is helpful to distinguish between two categories of stakeholders who will be interested in measurements of effectiveness:

- **Experts** are the researchers, designers, operators, and advanced users with appropriate scientific or professional credentials who have a technical understanding of the way in which a system works and are well-versed in evaluation methods and the results they generate.
- **Nonexperts** are the legislators, judges, lawyers, prosecutors, litigants, communities, victims, defendants, and system advocates whose work or legal outcomes may, even if only indirectly, be affected by the results

of a given system. These individuals, however, may not have a technical understanding of the way in which a system operates. Furthermore, they may have little experience in conducting scientific evaluations and interpreting their results.

Effectiveness metrics must meet the needs of *both* expert *and* nonexpert consumers.

- With respect to experts, the purpose of an effectiveness metric is *to advance both long-term research and more immediate product development, maintenance, and oversight*. To achieve that purpose, it is appropriate to define a fine-grained metric that may not be within the grasp of the nonexpert. Researchers and developers will be acting on the information provided by such a metric, so it should be tailored to their needs.
- With respect to nonexperts, including the general public, the purpose of an effectiveness metric is *to advance informed trust*, meaning trust that is based on sound evidence that the A/IS have met, or will meet, their intended objectives, taking into account both the immediate purpose and the contextual purpose of preserving and fostering important values such as human rights, dignity, and well-being. For this purpose, it will be necessary to define a metric that can serve as a readily understood summary measure of effectiveness. This metric must provide a simple, direct answer to the question of how effective a given system is. Automobile safety ratings are an example of this sort of metric. For automobile designers and engineers, the summary metrics are not sufficiently fine-grained to give immediately actionable information; for consumers, however, the metrics, insofar as they are accurate, empower them to make better-informed buying decisions.

For the purpose of fostering informed trust in A/IS adopted in the legal system, the most important goal is to establish a clear measure of effectiveness that can be understood by nonexperts. However, significant obstacles to achieving this goal include (a) developer incentives that prioritize research and development, along with the metrics that support such efforts, and (b) market forces that inhibit, or do not encourage, consumer-facing metrics. For those reasons, it is important that the selection and definition of the operative metrics draw on input not only from the A/IS creators but from other stakeholders as well; only under these conditions will a consensus form around the meaningfulness of the metrics.

What measurement practices foster informed trust?

By equipping both experts and nonexperts with accurate information regarding the capabilities and limitations of a given system, measurements of effectiveness can provide society with information needed to adopt and apply A/IS in a thoughtful, carefully considered, beneficial manner.[53]

In order for the practice of measuring effectiveness to realize its full potential for fostering trust and mitigating the risks of uninformed adoption and uninformed avoidance of adoption, it must have certain features:

Law

- **Meaningful metrics:** As noted above, an essential element of a measurement practice is a metric that provides an accurate and readily understood gauge of effectiveness. The metric should provide clear and actionable information as to the extent to which a given application has, or has not, met its objective so that potential users of the results of the application can respond accordingly. For example, in legal discovery, both recall and precision have done this well and have contributed to the acceptance of the use of A/IS for this purpose.[54]
- **Sound methods:** Measures of effectiveness must be obtained by scientifically sound methods. If, for example, measures are obtained by sampling, those sample-based estimates must be the result of sound statistical procedures that hold up to objective scrutiny.
- **Valid data:** Data on which evaluations of effectiveness are conducted should accurately represent the actual data to which the given A/IS would be applied and should be vetted for potential bias. Any data sets used for benchmarking or testing should be collected, maintained, and used in accordance with principles for the protection of individual privacy and agency.[55]
- **Awareness and consensus:** Measurement practices must not only be technically sound in terms of metrics, methods, and data, but they must also be widely understood and accepted as evidence of effectiveness.
- **Implementation:** Measurement practices must be both practically feasible and actually implemented, i.e., widely adopted by practitioners[56].
- **Transparency.** Measurement methods and results must be open to scrutiny by experts and the general public.[57] Without such scrutiny, the measurements will not be trusted and will be incapable of fulfilling their intended purpose.[58]

In seeking to advance informed trust in A/IS, policymakers should formulate policies and promote standards that encourage sound measurement practices, especially those that incorporate the key features.

Additional note. While in all circumstances all four principles discussed in this chapter (Effectiveness, Competence, Accountability, Transparency) will have something to contribute to the fostering of informed trust, it is not the case that in every circumstance all four principles will contribute equally to the fostering of trust. In some circumstances, a well-established measure of effectiveness, obtained by proven and accepted methods, may go a considerable way, on its own, in fostering trust in a given application—or distrust, if that is what the measurements indicate. In such circumstances, the challenges presented by the other principles, e.g., the challenge of adhering to the principle of transparency while respecting intellectual property considerations, may become of secondary importance.

Law

Illustration—Effectiveness

The search for factual evidence in large document collections in US civil or criminal proceedings has traditionally involved page-by-page manual review by attorneys. Starting in the 1990s, the proliferation of electronic data, such as email, rendered manual review prohibitively costly and time-consuming. By 2008, A/IS designed to substantially automate review of electronic data (a task known as "e-discovery") were available. Yet, adoption remained limited. Chief among the obstacles to adoption was a concern about the effectiveness, and hence defensibility in court, of A/IS in e-discovery. **Simply put, practitioners and courts needed a sound answer to a simple question: "Does it work?"**

Starting in 2006, the US NIST [59] conducted studies to assess that question.[60] The studies focused on, among others, two sound statistical metrics, both expressed as easy-to-understand percentages:[61,62]

- **Recall**, which is a gauge of the extent to which all the relevant documents were retrieved. For example, if there were 1,000 relevant documents to be found in the collection, and the review process identified 700 of them, then it achieved 70% recall.
- **Precision**, which is a gauge of the extent to which the documents identified as relevant by a process were actually relevant. For example, if for every two relevant documents the system captured, it also captured a nonrelevant one (i.e., a false positive), then it achieved 67% precision.

The studies provided empirical evidence that some systems could achieve high scores (80%) according to both metrics.[63] In a seminal follow-up study, Maura R. Grossman and Gordon V. Cormack found that two automated systems did, in fact, "conclusively" outperform human reviewers.[64] Drawing on the results of that study, Magistrate Judge Andrew Peck, in an opinion with far-reaching consequences, gave court approval for the use of A/IS to conduct legal discovery.[65]

The story of the TREC Legal Track's role in facilitating the adoption of A/IS for legal fact-finding contains a few lessons:

- **Metrics:** By focusing on recall and precision, the TREC studies quantified the effectiveness of the systems evaluated in a way that legal practitioners could readily understand.
- **Benchmarks:** The TREC studies filled an important gap: independent, scientifically sound evaluations of the effectiveness of A/IS applied to the real-world challenge of legal e-discovery.
- **Collaboration:** The founders of the TREC studies and the most successful participants came from both scientific and legal backgrounds, demonstrating the importance of multidisciplinary collaboration.

The TREC studies are a shining example of how the truth-seeking protocols of science can be used to advance the truth-seeking protocols of the law. They can serve as a conceptual basis for future benchmarking efforts, as well as the development of standards and certification programs to support informed trust when it comes to effectiveness of A/IS deployed in legal systems.[66]

Law

Recommendations

1. Governments should fund and support the establishment of ongoing benchmarking exercises designed to provide valid, publicly accessible measurements of the effectiveness of A/IS deployed, or potentially deployed, in the legal system. That support could take a number of forms, ranging from direct sponsorship and oversight—for example, by nonregulatory measurement laboratories such as the US NIST—to indirect support by the recognition of the results of a credible third-party benchmarking exercise for the purposes of meeting procurement and contracting requirements. All government efforts in this regard should be transparent and open to public scrutiny.
2. Governments should facilitate the creation of data sets that can be used for purposes of evaluating the effectiveness of A/IS as applied in the legal system. In assisting in the creation of such data sets, governments and administrative agencies will have to take into consideration potentially competing societal values, such as the protection of personal data, and arrive at solutions that maintain those values while enabling the creation of usable, real-world data sets. All government efforts in this regard should be transparent and open to public scrutiny.
3. Creators of A/IS to be applied to legal matters should pursue valid measures of the effectiveness of their systems, whether through participation in benchmarking exercises or through conducting single-system validation exercises. Creators should describe the procedures and results of the testing in clear language that is understandable to both experts and nonexperts, and should do so without disclosing intellectual property. Further, the descriptions should be open to examination by all stakeholders, including, when appropriate, the general public.
4. Researchers engaged in the study and development of A/IS for use in the legal system should seek to define meaningful metrics that gauge the effectiveness of the systems they study. In selecting and defining metrics, researchers should seek input from all stakeholders in the outcome of the given application of A/IS in the legal system. The metrics should be readily understandable by experts and nonexperts alike.
5. Governments and industry associations should undertake educational efforts to inform both those engaged in the operation of A/IS deployed in the legal system and those affected by the results of their operation of the salient measures of effectiveness and what they can indicate about the capabilities and limitations of the A/IS in question.
6. Creators of A/IS for use in the legal system should ensure that the effectiveness metrics defined by the research community are readily obtainable and accessible to all stakeholders, including, when appropriate, the general public. Creators should provide guidance on how to interpret and respond to the metrics generated by the system.
7. Operators of A/IS applied to a legal task should follow the guidance on the measurement of effectiveness provided for

the A/IS being used. This includes guidance about which metrics to obtain, how and when to obtain them, how to respond to given results, when it may be appropriate to follow alternative methods of gauging effectiveness, and so on.

8. In interpreting and responding to measurements of the effectiveness of A/IS applied to legal problems or questions, allowance should be made by those interpreting the results for variation in the specific objectives and circumstances of a given deployment of A/IS. Quantitative results should be supplemented by qualitative evaluation of the practical significance of a given outcome and whether it indicates a need for remediation. This evaluation should be done by an individual with the technical expertise and pragmatic experience needed to make a sound judgment.
9. Industry associations or other organizations should collaborate on developing standards for measuring and reporting on the effectiveness of A/IS. These standards should be developed with input from both the scientific and legal communities.
10. Recommendation 1 under Issue 2, with respect to effectiveness.
11. Recommendation 2 under Issue 2, with respect to effectiveness.

Further Resources

- *Da Silva Moore v. Publicis Groupe*, 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012).
- C. Garvie, A. M. Bedoya, and J. Frankle, "The Perpetual Line-Up: Unregulated Police Face Recognition in America," Georgetown Law, Center on Privacy & Technology, Oct. 2016.
- M. R. Grossman 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, vol. 17, no. 3, 2011.
- B. Hedin, D. Brassil, and A. Jones, "On the Place of Measurement in E-Discovery," in Perspectives on Predictive Coding and Other Advanced Search Methods for the Legal Practitioner, J. R. Baron, R. C. Losey, and M. D. Berman, Eds. Chicago: American Bar Association, 2016.
- J. A. Kroll, "The fallacy of inscrutability," Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences, vol. 376, no. 2133, Oct. 2018.
- D. W. Oard, J. R. Baron, B. Hedin, D. Lewis, and S. Tomlinson, "Evaluation of Information Retrieval for E-Discovery," Artificial Intelligence and Law, vol. 18, no. 4, pp. 347-386, Aug. 2010.
- The Sedona Conference, "The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process," The Sedona Conference Journal, vol. 15, pp. 265-304, 2014.
- M. T. Stevenson, "Assessing Risk Assessment in Action," Minnesota Law Review, vol. 103, June 2018.

- "Global Governance of AI Roundtable: Summary Report 2018," World Government Summit, 2018.
- High-Level Expert Group on Artificial Intelligence, "DRAFT Ethics Guidelines for Trustworthy AI: Working Document for Stakeholders' Consultation," The European Commission. Brussels, Belgium: Dec. 18, 2018.

Issue 4: Competence

How can specification of the knowledge and skills required of the human operator(s) of A/IS foster informed in the suitability of A/IS for adoption in legal systems?

Background

An essential component of informed trust in a technological system, especially one that may affect us in profound ways, is confidence in the competence of the operator(s) of the technology. We trust surgeons or pilots with our lives because we have confidence that they have the knowledge, skills, and experience to apply the tools and methods needed to carry out their tasks effectively. We have that confidence because we know that these operators have met rigorous professional and scientific accreditation standards before being allowed to step into the operating room or cockpit. This informed trust in operator competence is what gives us confidence that surgery or air travel will result in the desired outcome. No such standards of operator competence currently exist with respect to A/IS applied in legal systems, where the life, liberty, and rights of citizens can be at stake. That absence of standards hinders the trustworthy adoption of A/IS in the legal domain.

The human operator is an integral component of A/IS

Almost all current applications of A/IS in legal systems, like those in most other fields, require human mediation and likely will continue to do so for the near future. This human mediation, post design and post development, will take a number of forms, including decisions about (a) whether or not to use A/IS for a given purpose,[67] (b) the data used to train the systems, (c) settings for system parameters to be used in generating results, (d) methods of validating results, (e) interpretation and application of the results, and so on. Because these systems' outcomes are a function of all their components, including the human operator(s), their effectiveness, and by extension trustworthiness, will depend on their human operator(s).

Despite this, there are few standards that specify how humans should mediate applications of A/IS in legal systems, or what knowledge qualifies a person to apply A/IS and interpret their results.[68] This reality is especially troubling for the instances in which the life, rights, or liberty of humans are at stake. Today, while professional codes of ethics for lawyers are beginning to include among their

Law

requirements an awareness and understanding of technologies with legal application,[69] the operators of A/IS in legal systems are essentially deemed to be capable of determining their own competence: lawyers or IT professionals operating in civil discovery, correctional officers using risk assessment algorithms, and law enforcement agencies engaging in predictive policing or using automated surveillance technologies. All are mostly able to use A/IS without demonstrating that they understand the operation of the system they are using or that they have any particular set of consensus competencies.[70]

The lack of competency requirements or standards undermines the establishment of informed trust in the use of A/IS in legal systems. If courts, legal practitioners, law enforcement agencies, and the general public are to rely on the results of A/IS when applied to tasks traditionally carried out by legal professionals, they must have grounds for believing that those operating A/IS will possess the requisite knowledge and skill to understand the conditions and methods for operating the systems effectively, including evaluating the data on which the A/IS trained, the data to which they are applied, the results they produce, and the methods and results of measuring the effectiveness the systems. Applied incompetently, A/IS could produce the opposite intended effect. Instead of improving a legal system—and bringing about the gains in well-being that follow from such improvements—they may undermine both the fairness and effectiveness of a legal system and trust in its fairness and effectiveness, creating conditions for social disorder and the deterioration of human well-being that would follow from that disorder. By way of illustration:

- A city council might misallocate funds for policing across city neighborhoods because it relies on the output of an algorithm that directs attention to neighborhoods based on arrest rates rather than actual crime rates.[71]
- In civil justice, A/IS applied in a search of documents to uncover relevant facts may fail to do so because an operator without sufficient competence in statistics may materially overestimate the accuracy of the system, thus ceasing vital fact-finding activities.[72]
- In the money bail system, reliance on A/IS to reduce bias may instead perpetuate it. For example, if a judge does not understand whether an algorithm makes sufficient contextual distinctions between gradations of offenses,[73] that judge would not able to probe the output of the A/IS and make a well-informed use of it.
- In the criminal justice system, an operator using A/IS in sentencing decision-support may fail to identify bias, or to assess the risk of bias, in the results generated by the A/IS,[74] unfairly depriving a citizen of his or her liberty or prematurely granting an offender's release, increasing the risk of recidivism.

More generally, without the confidence that A/IS operators will apply the technology as intended and supervise it appropriately, the general public will harbor fear, uncertainty, and doubt about the use of A/IS in legal systems and potentially about the legal systems themselves.

Law

Fostering informed trust in the competence of human operators

If negative outcomes such as those just described are to be avoided, **it will be necessary to include among norms for the adoption of A/IS in a legal system a provision for building informed trust in the operators of A/IS**. Building trust will require articulating standards and best practices for two groups of agents involved in the deployment of A/IS: creators and operators.

On the one hand, those engaged in the design, development, and marketing of A/IS must commit to specifying the knowledge, skills, and conditions required for the safe, ethical, and effective deployment and operation of the systems.[75] On the other hand, those engaged in actually operating the systems, including both legal professionals and experts acting in the service of legal professionals, must commit to adhering to these requirements in a manner consistent with other operative legal, ethical, and professional requirements. The precise nature of the competency requirements will vary with the nature and purpose of the A/IS and what is at stake in their effective operation. The requirements for the operation of A/IS designed to assist in the creation of contracts, for example, might be less stringent than those for the operation of A/IS designed to assess flight risk, which could affect the liberty of individual citizens.

A corollary of these provisions is that education and training in the requisite skills should be available and accessible to those who would operate A/IS, whether that training is provided through professional schools, such as law school; through institutions providing ongoing professional training, such as, for federal judges in the United States, the Federal Judicial Center; through professional and industry associations, such as the American Bar Association; or through resources accessible by the general public.[76] Making sure such training is available and accessible will be essential to ensuring that the resources needed for the competent operation of A/IS are widely and equitably distributed.[77]

It will take a combined effort of both creators and operators to ensure both that A/IS designed for use in legal systems are properly applied and that those with a stake in the effective functioning of legal systems—including legal professionals, of course, but also decision subjects, victims of crime, communities, and the general public—will have informed trust, or, for that matter, informed distrust (if that is what a competence assessment finds) in the competence of the operators of A/IS as applied to legal problems and questions.[78]

Illustration—Competence

Included among the offerings of Amazon Web Services is an image and video analysis service known as Amazon Rekognition.[79] The service is designed to enable the recognition of text, objects, people, and actions in images and videos. The technology also enables the search and comparison of faces, a feature with potential law enforcement and national security applications, such as comparing faces identified in video taken by a security camera with those in a database of jail booking photos. Attracted by

Law

the latter feature, police departments in Oregon and Florida have undertaken pilots of Rekognition as a tool in their law enforcement efforts.[80]

In 2018, the American Civil Liberties Union (ACLU), a frequent critic of the use of facial recognition technologies by law enforcement agencies,[81] conducted a test of Rekognition. The test consisted of first constructing a database of 25,000 booking photos ("mugshots") then comparing publicly available photos of all then-current members of the US Congress against the images in the database. The test found that Rekognition incorrectly matched the faces of 28 members of Congress with faces of individuals who had been arrested for a crime.[82] The ACLU argues that the high number of false positives generated by the technology shows that police use of facial recognition technologies generally (and of Rekognition in particular) poses a risk to the privacy and liberty of law-abiding citizens. The ACLU has used the results of its test of Rekognition to support its proposal that Congress enact a moratorium on the use of facial recognition technologies by law enforcement agencies until stronger safeguards against their misuse, and potential abuse, can be put in place.[83]

In response to the ACLU report, Amazon noted that the ACLU researchers, in conducting their study, had applied the technology utilizing a similarity threshold (a gauge of the likelihood of a true match) of 80%, a threshold that casts a fairly wide net for potential matches (and hence generates a high number of false positives). For applications in which there are greater costs associated with false positives (e.g., policing), Amazon recommends utilizing a similarity threshold value of 99% or above to reduce accidental misidentification.[84] Amazon also noted that, in all law enforcement use cases, it would be expected that the results of the technology would be reviewed by a human before any actual police action would be undertaken.

The story of the ACLU's testing of Rekognition and Amazon's response to the test highlights the importance of specifying and adhering to guidelines for competent use.[85] Had a law enforcement agency used the technology in the way it was used in the ACLU test, it would, in most legitimate use cases, be guilty of incompetent use. At the same time, Amazon is not free of blame insofar as it did not specify prominently and clearly the competency guidelines for effective use of the technology in support of law enforcement efforts, as well as the risks that might be incurred if those guidelines are not followed. Competent use[86] follows both from the A/IS creator's specification of well-grounded[87] competency guidelines and from the A/IS operator's adherence to those guidelines.[88]

Recommendations

1. Creators of A/IS for application in legal systems should provide clear and accessible guidance for the knowledge, skills, and experience required of the human operators of the A/IS if the systems are to achieve expected levels of effectiveness. Included in that guidance should be a delineation of the risks involved if those requirements are not met. Such guidance should be

Law

documented in a form that is accessible and understandable by both experts and the general public.

2. Creators and developers of A/IS for application in legal systems should create written policies that govern how the A/IS should be operated. In creating these policies, creators and developers should draw on input from the legal professionals who will be using the A/IS they are creating. The policies should include:
   - the specification of the real-world applications for the A/IS;
   - the preconditions for their effective use;
   - the training and skills that are required for operators of the systems;
   - the procedures for gauging the effectiveness of the A/IS;
   - the considerations to take into account in interpreting the results of the A/IS;
   - the outcomes that can be expected by both operators and other affected parties when the A/IS are operated properly; and
   - the specific risks that follow from improper use.

   The policies should also specify circumstances in which it might be necessary for the operator to override the A/IS. All such policies should be publicly accessible.

3. Creators and developers of A/IS to be applied in legal systems should integrate safeguards against the incompetent operation of their systems. Safeguards could include issuing notifications and warnings to operators in certain conditions, requiring, as appropriate, acknowledgment of receipt; limiting access to A/IS functionality based on the operator's level of expertise; enabling system shut-down in potentially high-risk conditions; and more. These safeguards should be flexible and governed by context-sensitive policies set by competent personnel of the entity (e.g., the judiciary), utilizing the A/IS to address a legal problem.

4. Governments should provide that any individual whose legal outcome is affected by the application of A/IS should be notified of the role played by A/IS in that outcome. Further, the affected party should have recourse to appeal to the judgment of a competent human being.

5. Professionals engaged in the creation, practice, interpretation, and enforcement of the law, such as lawyers, judges, and law enforcement officers, should recognize the specialized scientific and professional expertise required for the ethical and effective application of A/IS to their professional duties. The professional associations to which such legal practitioners belong, such as the American Bar Association, should, through both educational programs and professional codes of ethics, seek to ensure that their members are well informed about the scientific and technical competency requirements for the effective and trustworthy application of A/IS to the law.[89]

6. The operators of A/IS applied in legal systems—whether the operator is a specialist in A/IS or a legal professional—should

understand the competencies required for the effective performance of their roles and should either acquire those competencies or identify individuals with those competencies who can support them in the performance of their roles. The operator does not need to be an expert in all the pertinent domains but should have access to individuals with the requisite expertise.

7. Recommendation 1 under Issue 2, with respect to competence.
8. Recommendation 2 under Issue 2, with respect to competence.

Further Resources

- C. Garvie, A. M. Bedoya, and J. Frankle, "The Perpetual Line-Up: Unregulated Police Face Recognition in America," Georgetown Law, Center on Privacy & Technology, Oct. 2016.
- International Organization for Standardization, *ISO/IEC 27050-3: Information technology—Security techniques—Electronic discovery—Part 3: Code of practice for electronic discovery*, Geneva, 2017.
- J. A. Kroll, "The fallacy of inscrutability," Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences, vol. 376, no. 2133, Oct. 2018.
- A. G. Ferguson, "Policing Predictive Policing," Washington University Law Review, vol. 94, no. 5 2017.
- "Global Governance of AI Roundtable: Summary Report 2018," World Government Summit, 2018.

Issue 5: Accountability

How can the ability to apportion responsibility for the outcome of the application of A/IS foster informed trust in the suitability of A/IS for adoption in legal systems?

Background

Apportioning responsibility. An essential component of informed trust in a technological system is confidence that it is possible, if the need arises, to apportion responsibility among the human agents engaged along the path of its creation and application: from design through to development, procurement, deployment,[90] operation, and, finally, validation of effectiveness. Unless there are mechanisms to hold the agents engaged in these steps accountable, it will be difficult or impossible to assess responsibility for the outcome of the system under any framework, whether a formal legal framework or a less formal normative framework. A model of A/IS creation and use that does not have such mechanisms will also lack important forms of deterrence against poorly thought-out design, casual adoption, and inappropriate use of A/IS.

Simply put, a system that produces outcomes for which no one is responsible cannot be trusted. Those engaged in creating, procuring, deploying, and operating such a system will lack the discipline engendered by the clear

assignment of responsibility. Meanwhile, those affected by the results of the system's operation will find their questions around a given result inadequately answered, and errors generated by the system will go uncorrected. In the case of A/IS applied in a legal system, where an individual's basic human rights may be at issue, these questions and errors are of fundamental importance. In such circumstances, the only options are either blind trust or blind distrust. Neither of those options is satisfactory, especially in the case of a technological system applied in a domain as fundamental to the social order as the law.

Challenges to accountability

In the case of A/IS, whether applied in a legal system or another domain, maintaining accountability can be a particularly steep challenge. This challenge to accountability is because of both the perceived "black box" nature of A/IS and the diffusion of responsibility it brings.

The perception of A/IS as a black box stems from the opacity that is an inevitable characteristic of a system that is a complex nexus of algorithms, computer code, and input data. As observed by Joshua New and Daniel Castro of the Information Technology and Innovation Foundation:

The most common criticism of algorithmic decision-making is that it is a "black box" of extraordinarily complex underlying decision models involving millions of data points and thousands of lines of code. Moreover, the model can change over time, particularly when using machine learning algorithms that adjust the model as the algorithm encounters new data.[91]

This opacity of the systems makes it challenging to trace cause to effect,[92] which, in turn, makes it difficult or even impossible, to draw lines of responsibility.

The diffuseness challenge stems from the fact that even the most seemingly straightforward A/IS can be complex, with a wide range of agents—systems designers, engineers, data analysts, quality control specialists, operators, and others—involved in design, development, and deployment. Moreover, some of these agents may not even have been engaged in the development of the A/IS in question; they may have, for example, developed open-source components that were intended for an entirely different purpose but that were subsequently incorporated into the A/IS. This diffuseness of responsibility poses a challenge to the maintenance of accountability.[93] As Matthew Scherer, a frequent writer and speaker on topics at the intersection of law and A/IS, observes:

The sheer number of individuals and firms that may participate in the design, modification, and incorporation of an AI system's components will make it difficult to identify the most responsible party or parties. Some components may have been designed years before the AI project had even been conceived, and the components' designers may never have envisioned, much less intended, that their designs would be incorporated into any AI system, still less the specific AI system that caused harm. In such circumstances, it may seem unfair to assign

blame to the designer of a component whose work was far-removed in both time and geographic location from the completion and operation of the AI system.[94]

Examples include the following:

- When a judge's ruling includes a long prison sentence, based in part on a flawed A/IS-enabled process that erroneously deemed a particular person to be at high risk of recidivism, who is responsible for the error? Is it the A/IS designer, the person who chose the data or weighed the inputs, the prosecution team who developed and delivered the risk profile to the court, or the judge who did not have the competence to ask the appropriate questions that would have enabled a clearer understanding of the limitations of the system? Or is responsibility somehow distributed among these various agents?[95]
- When a lawyer engaged in civil or criminal discovery believes, erroneously, that all the relevant information was found when using A/IS in a data-intensive matter, who is responsible for the failure to gather important facts? The A/IS designer who typically would have had no ability to foretell the specific circumstances of a given matter, the legal or IT professional who operated the A/IS or erroneously measured its effectiveness, or the lawyer who made a representation to his or her client, to the court, or to investigatory agencies?
- When a law enforcement officer, relying on A/IS, erroneously identifies an individual as being more likely to commit a crime than another, who is responsible for the resulting encroachment on the civil rights of the person erroneously targeted? Is it the A/IS designer, the individual who selected the data on which to train the algorithm, the individual who chose how the effectiveness of the A/IS would be measured,[96] the experts who provided training to the officer, or the officer himself or herself?

As a result of the challenges presented by the opacity and diffuseness of responsibility in A/IS, the present-day answer to the question, "Who is accountable?" is, in far too many instances, "It's hard to say." This is a response that, in practice, means "no one" or, equally unhelpful, "everyone". Such failure to maintain accountability will undermine efforts to bring A/IS (and all their potential benefits) into legal systems based on informed trust.

Maintaining accountability and trust in A/IS

Although maintaining accountability in complex systems can be a challenge, it is one that must be met in order to engender informed trust in the use of A/IS in the legal domain. "Blaming the algorithm" is not a substitute for taking on the challenge of maintaining transparent lines of responsibility and establishing norms of accountability.[97] This is true even if we allow that, given the complexity of the systems in question, some number of "systems accidents" is inevitable.[98] Informed trust in a system does not require a belief that zero errors will occur; however, it does require a belief that there are mechanisms in place for addressing errors when

Law

they do occur. Accountability is an essential component of those mechanisms.

In meeting the challenge, it should be recognized that there are existing norms and controls that have a role to play in ensuring that accountability is maintained. For example, contractual arrangements between the A/IS provider and a party acquiring and applying a system may help to specify who is (and is not) to be held liable in the event the system produces undesirable results. Professional codes of ethics may also go some way toward specifying the extent to which lawyers, for example, are responsible for the results generated by the technologies they use, whether they operate them directly or retain someone else to do so. Judicial systems may have procedures for assessing responsibility when a citizen's rights are improperly infringed. As illustrated by the cases described above, however, existing norms and controls, while helpful, are insufficient in themselves to meet the specific challenge represented by the opacity and diffuseness of A/IS. To meet the challenge further steps must be taken.[99]

The first step is ensuring that all those engaged in the creation, procurement, deployment, operation, and testing of A/IS recognize that, if accountability is not maintained, these systems will not be trusted. In the interest of maintaining accountability, these stakeholders should take steps to clarify lines of responsibility throughout this continuum, and make those lines of responsibility, when appropriate, accessible to meaningful inquiry and audit.

The goal of clarifying lines of responsibility in the operation of A/IS is to implement a governing model that specifies who is responsible for what, and who has recourse to which corrective actions, i.e., a trustworthy model that ensures that it will admit actionable answers should questions of accountability arise. Arriving at an effective model will require the participation of those engaged in the creation and operation of A/IS, those affected by the results of their use, and those with the expertise to understand how such a model would be used in a given legal system. For example:

- Individuals responsible for the design of A/IS will have to maintain a transparent record of the sources of the various components of their systems, including identification of which components were developed in-house and which were acquired from outside sources, whether open source or acquired from another firm.
- Individuals responsible for the design of A/IS will have to specify the roles, responsibilities, and potential subsequent liabilities of those who will be engaged in the operation of the systems they create.
- Individuals responsible for the operation of a system will have to understand their roles, responsibilities, potential liabilities, and will have to maintain documentation of their adherence to requirements.
- Individuals affected by the results of the operation of A/IS, e.g., a defendant in a criminal proceeding, will have to be given access to information about the roles and responsibilities of those involved in relevant

Law

aspects of the creation, operation, and validation of the effectiveness of the A/IS affecting them.[100]

- Individuals with legal and political training (e.g., jurists, regulators, as well as legal and political scholars) will have to ensure that any model that is created will provide information that is in fact actionable within the operative legal system.

A governing model of accountability that reflects the interests of all these stakeholders will be more effective both at deterring irresponsible design or use of A/IS before it happens and at apportioning responsibility for an undesirable outcome when it does happen.[101]

Pulling together the input from the various stakeholders will likely not take place without some amount of institutional initiative. Organizations that employ A/IS for accomplishing legal tasks—private firms, regulatory agencies, law enforcement agencies, judicial institutions—should therefore develop and implement policies that will advance the goal of clarifying lines of responsibility. Such policies could take the form of, for example, designating an official specifically charged with oversight of the organization's procurement, deployment, and evaluation of A/IS as well as the organization's efforts to educate people both inside and outside the organization on its use of A/IS. Such policies might also include the establishment of a review board to assess the organization's use of A/IS and to ensure that lines of responsibility for the outcomes of its use are maintained. In the case of agencies, such as police departments, whose use of A/IS could impact the general public, such review boards would, in the interest of legitimacy, have to include participation from various citizens' groups, such as those representing defendants in the criminal system as well as those representing victims of crime.[102]

The goal of opening lines of responsibility to meaningful inquiry is to ensure that an investigation into the use of A/IS will be able to isolate responsibility for errors (or potential errors) generated by the systems and their operation.[103] This means that all those engaged in the design, development, procurement, deployment, operation, and validation of the effectiveness of A/IS, as well as the organizations that employ them, must in good faith be willing to participate in an audit, whether the audit is a formal legal investigation or a less formal inquiry. They must also be willing to create and preserve documentation of key procedures, decisions, certifications,[104] and tests made in the course of developing and deploying the A/IS.[105]

The combination of a governing model of accountability and an openness to meaningful audit will allow the maintenance of accountability, even in complex deployments of A/IS in the service of a legal system.

Additional note 1. The principle of accountability is closely linked with each of the other principles intended to foster informed trust in A/IS: effectiveness, competence, and transparency. With respect to effectiveness, evidence of attaining key metrics and benchmarks to confirm that A/IS are functioning as intended may put questions of where, among creators,

owners, and operators, responsibility for the outcome of a system lies on a sound empirical footing. With respect to competence, operator credentialing and specified system handoffs enable a clear chain of responsibility in the deployment of A/IS.[106] With respect to transparency, providing a view into the general design and methods of A/IS, or even a specific explanation for a given outcome, can help to advance accountability.

Additional note 2. Closely related to accountability is the trust that follows from knowing that a human expert is guiding the A/IS and is capable of overriding them, if necessary. Subjecting humans to automated decisions not only raises legal and ethical concerns, both from a data protection[107] and fundamental rights perspective,[108] but also will likely be viewed with distrust if the human component, which can introduce circumstantial flexibility in the interest of realizing an ethically superior outcome, is missing. In addition to ensuring technical safety and reliability of A/IS used in the course of decision-making processes, the legal system should also, where appropriate, provide for the possibility of an appeal for review by a human judge. Careful attention must be paid to the design of corresponding appeal procedures.[109]

Illustration—Accountability

Over the last two decades, criminal justice agencies have increasingly embraced predictive tools to assist in the determination for bail, sentencing, and parole. A mix of companies, government agencies, nonprofits, and universities have built and promoted tools that provide a likelihood that someone may fail to appear or may commit a new crime or a new violent act. While math has played a role in these determinations since at least the 1920s,[110] a new interest in accountability and transparency has brought novel legal challenges to these tools.

In 2013, Eric Loomis was arrested for a drive-by shooting in La Crosse, Wisconsin. No one was hit, but Loomis faced prison time. Loomis denied involvement in the shooting, but waived his right to trial and entered a guilty plea to two of the less severe offenses with which he was charged: attempting to flee a traffic officer and operating a motor vehicle without the owner's consent. The judge sentenced him to six years in prison, saying he was "high risk". The judge based this conclusion, in part, on the risk assessment score given by Compas, a secret and privately held algorithmic tool used routinely by the Wisconsin Department of Corrections.

On appeal, Loomis made three major arguments, two focused on accountability.[111] First, the tool's proprietary nature—the underlying code was not made available to the defense—made it impossible to test its scientific validity. Second, the tool inappropriately considered gender in making its determination.

A unanimous Wisconsin Supreme Court ruled against Loomis on both arguments.

The court reasoned that knowing the inputs and output of the tool, and having access to validating studies of the tool's accuracy, were sufficient to prevent infringement of Loomis' due process.[112] Regarding the use of gender—a protected class in the United States—the court said he did not show that there was a reliance on gender in making the output or sentencing decision.

Law

Without the ability to interrogate the tool and know how gender is used, the court created a paradox with its opinion.

The *Loomis* decision represents the challenges that judges have balancing accountability of "black boxed" A/IS and trade secret protections.[113] Other decisions have sided against accountability of other risk assessments,[114] probabilistic DNA analysis tools,[115] and government remote hacking investigation software.[116] Siding with accountability, a federal judge found that the underlying code of a probability software used in DNA comparisons was admissible and relevant to a pretrial hearing where the admissibility of expert testimony is challenged.[117]

These issues will continue to be litigated as A/IS tools continue to proliferate in judicial systems. To that end, as the *Loomis* court notes, "The justice system must keep up with the research and continuously assess the use of these tools."

Recommendations

1. Creators of A/IS to be applied in a legal system should articulate and document well-defined lines of responsibility, among all those who would be engaged in the development and operation of the A/IS, for the outcome of the A/IS.
2. Those engaged in the adoption and operation of A/IS to be applied in a legal system should understand their specific responsibilities for the outcome of the A/IS as well as their potential liability should the A/IS produce an outcome other than that intended. In the case of A/IS, many questions of legal liability remain unsettled. Adopters and operators of A/IS should nevertheless understand to what extent they could, *potentially*, be held liable for an undesirable outcome.
3. When negotiating contracts for the provision of A/IS products and services for use in the legal system, providers and buyers of A/IS should include contractual terms specifying clear lines of responsibility for the outcomes of the systems being acquired.
4. Creators and operators of A/IS applied in a legal system, and the organizations that employ them, should be amenable to internal oversight mechanisms and inquiries (or audits) that have the objective of allocating responsibility for the outcomes generated by the A/IS. In the case of A/IS adopted and deployed by organizations that have direct public interaction (e.g., a law enforcement agency), oversight and inquiry could also be conducted by external review boards. Being prepared for such inquiries means maintaining clear documentation of all salient procedures followed, decisions made, and tests conducted in the course of developing and applying the A/IS.
5. Organizations engaged in the development and operation of A/IS for legal tasks should consider mechanisms that will create individual and collective incentives for ensuring both that the outcomes of the A/IS adhere to ethical standards and that accountability for those outcomes is maintained, e.g., mechanisms to ensure that speed and efficiency are not rewarded at the expense of a loss of accountability.

6. Those conducting inquiries to determine responsibility for the outcomes of A/IS applied in a legal system should take into consideration all human agents involved in the design, development, procurement, deployment, operation, and validation of effectiveness of the A/IS and should assign responsibility accordingly.
7. Recommendation 1 under Issue 2, with respect to accountability.
8. Recommendation 2 under Issue 2, with respect to accountability.

Further Resources

- N. Diakopoulos, S. Friedler, M. Arenas, S. Barocas, M. Hay, B. Howe, H. V. Jagadish, K. Unsworth, A. Sahuguet, S. Venkatasubramanian, C. Wilson, C. Yu, and B. Zevenbergen, "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms," FAT/ML.
- F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. J. Gershman, D. O'Brien, S. Shieber, J. Waldo, D. Weinberger, and A. Wood, "Accountability of AI Under the Law: The Role of Explanation," Berkman Center Research Publication Forthcoming; Harvard Public Law Working Paper, no. 18-07, Nov. 3, 2017.
- European Commission for the Efficiency of Justice. *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment*. Strasbourg, 2018.
- J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, "Accountable Algorithms," University of Pennsylvania Law Review, vol. 165, pp. 633-705. Feb. 2017.
- J. New and D. Castro, "How Policymakers Can Foster Algorithmic Accountability," Information Technology and Innovation Foundation, May 21, 2018.
- M. U. Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," Harvard Journal of Law & Technology, vol. 29. no. 2, pp. 369-373, 2016.
- J. Tashea, "Calculating Crime: Attorneys are Challenging the Use of Algorithms to Help Determine Bail, Sentencing and Parole," ABA Journal, March 2017.

Issue 6: Transparency

How can sharing information that explains how A/IS reached given decisions or outcomes foster informed trust in the suitability of A/IS for adoption in legal systems?

Background

Access to meaningful information.
An essential component of informed trust in a technological system is confidence that the information required for a human to understand why the system behaves a certain way in a specific circumstance (or would behave in

a hypothetical circumstance) will be accessible. Without transparency, there is no basis for trusting that a given decision or outcome of the system can be explained, replicated, or, if necessary, corrected.[118] Without transparency, there is no basis for informed trust that the system can be operated in a way that achieves its ends reliably and consistently or that the system will not be used in a way that impinges on human rights. In the case of A/IS applied in a legal system, such a lack of trust could undermine the credibility of the legal system itself.

Transparency and trust

Transparency, by prioritizing access to information about the operation and effectiveness of A/IS, serves the purpose of fostering informed trust in the systems. More specifically, transparency fosters trust that:

- the operation of A/IS and the results they produce are explainable;
- the operation and results of A/IS are fair;[119]
- the operation and results of A/IS are unbiased;
- the A/IS meet normative standards for operation and results;
- the A/IS are effective;
- the results of A/IS are replicable;[120] and
- those engaged in the design, development, procurement, deployment, operation, and validation of the effectiveness of A/IS can be held accountable, where appropriate, for negative outcomes, and that corrective or punitive action can be taken when warranted.

For A/IS used in a legal system to achieve their intended purposes, all those with a stake in the effective functioning of the legal system must have a well-grounded trust that the A/IS can meet these requirements. This trust can be fostered by transparency.

The elements of transparency

Transparency of A/IS in legal matters requires disclosing information about the design and operation of the A/IS to various stakeholders. In implementing the principle, however, we must, in the interest of both feasibility and effectiveness, be more precise both about the categories of stakeholders to whom the information will be disclosed, and about the categories of information that will be disclosed to those stakeholders.

Relevant stakeholders in a legal system include those who:

- operate A/IS for the purpose of carrying out tasks in civil justice, criminal justice, and law enforcement, such as a law enforcement officer who uses facial recognition tools to identify potential suspects;
- rely on the results of A/IS to make important decisions, such as a judge who draws on the results of an algorithmic assessment of recidivism risk in deciding on a sentence;
- are directly affected by the use of A/IS—a "decision subject", such as a defendant in a criminal proceeding whose bail terms are influenced by an algorithmic assessment of flight risk;

Law

- are indirectly affected by the results of A/IS, such as the members of a community that receives more or less police attention because of the results of predictive policing technology; and
- have an interest in the effective functioning of the legal system, such as judges, lawyers, and the general public.

Different types of relevant information can be grouped into high-level categories. As illustrated below, a taxonomy of such high-level categories may, for example, distinguish between:

- nontechnical procedural information regarding the employment and development of a given application of A/IS;
- information regarding data involved in the development, training, and operation of the system;
- information concerning a system's effectiveness/performance;
- information about the formal models that the system relies on; and
- information that serves to explain a system's general logic or specific outputs.

These more granular distinctions matter because different sorts of inquiries will require different sorts of information, and it is important to match the information provided to the actual needs of the inquiry. For example, an inquiry into a predictive policing system that misdirected police resources may not be much advanced by information about the formal models on which the system relied, but it may well be advanced by an explanation for the specific outcome.

On the other hand, an inquiry, undertaken by a designer or operator, into ways to improve system performance may benefit from access to information about the formal models on which the system relies.[121]

These distinctions also matter because there may be circumstances in which it would be desirable to limit access to a given type of information to certain stakeholders. For example, there may be circumstances in which one would want to identify an agent to serve as a public interest steward. For auditing purposes, this individual would have access to certain types of sensitive information unavailable to others. Such restrictions on information access are necessary if the transparency principle is not to impinge on other societal values and goals, such as security, privacy, and appropriate protection of intellectual property.[122]

The salience of the question, "*Who is given access to what information?*" is illustrated by Sentiment Meter, a technology developed by Elucd, a GovTech company that provides cities with near real-time understanding of how citizens feel about their government, in conjunction with the New York Police Department, to assist the NYPD in gauging citizens' views regarding police activity in their communities.[123] One of the stated goals of the program is to build public trust in the police department. In the interest of trust, should "the public" have access to all potentially relevant information, including how the system was designed and developed, what the input data are, who operates the system and what their qualifications are, how the system's effectiveness was tested, and why the public was not brought

into the process of construction? If the answer is that the general public should not have access to all this information, then who should? How do we define "the public?" Is it the whole community represented in its elected officials? Or should certain communities have greater access, for example, those most affected by controversial police practices such as stop, question, and frisk? Such questions must be answered if the program is to achieve its stated goals.

Transparency in practice

As just noted, although transparency can foster informed trust in A/IS applied in a legal system, **its practical implementation requires careful thought.** Requiring public access to all information pertaining to the operation and results of A/IS is neither necessary nor feasible. What is required is a careful consideration of who needs access to what information for the specific purpose of building informed trust. The following table is an example of a tool that might be used to match type of information to type of information consumer for the purpose of fostering trust.[124]

Law

Types of information that should be considered in determining transparency demands in relation to a given A/IS		Stakeholders whose interest in access to different types of information should be considered in determining the transparency demands in relation to a given application of A/IS			
High-level category	Specific type of information (examples) Disclosure of...	Operators	Decision-subjects	Public interest steward	General public
Procedural aspects regarding A/IS employment and development	the fact that a given context involves the employment of A/IS	N/A	?	?	?
	how the employment of the system was authorized	?	?	?	?
	who developed the system	?	?	?	?
	...				
Data involved in A/IS development and operation	the origins of training data and data involved in the operation of the system	?	?	?	?
	the kinds of quality checks that data was subject to and their results	?	?	?	?
	how data labels are defined and to what extent data involves proxy variables	?	?	?	?
	relevant data sets themselves	?	?	?	?
	...				
Effectiveness/ performance	the kinds of effectiveness/performance measurement that have occurred	?	?	?	?
	measurement results	?	?	?	?
	any independent auditing or certification	?	?	?	?
	...				
Model specification	the input variables involved	?	?	?	?
	the variable(s) that the model optimizes for	?	?	?	?
	tthe complete model (complete formal representation, source code, etc.)	?	?	?	?
	...				
Explanation	information concerning the system's general logic or functioning	?	?	?	?
	information concerning the determinants of a particular output [125]	?	?	?	?
	...				

Law

When it comes to deciding whether a specific type of information should be made available and, if so, which types of stakeholders should have access to it, there are various considerations, for example:

- The release of certain types of information may conflict with data privacy concerns, commercial or public policy interests—such as the promotion of innovation through appropriate intellectual property protections—and security interests, e.g., concerns about gaming and adversarial attacks. At the same time, such competing interests should not be permitted to be used, without specific justification, as a blanket cover for not adhering to due process, transparency, or accountability standards. The tension between these interests is particularly acute in the case of A/IS applied in a legal system, where the dignity, security, and liberty of individuals are at stake.[126]
- There is tension between the specific goal of explainability, which may argue for limits on system complexity, and system performance, which may be served by greater complexity, to the detriment of explainability.[127]
- One must carefully consider the question that is being asked in an inquiry into A/IS and what information transparency can actually produce to answer that question. Disclosure of A/IS algorithms or training data is, itself, insufficient to enable an auditor to determine whether the system was effective in a specific circumstance.[128] By analogy, transparency into drug manufacturing processes does not, itself, provide information about the actual effectiveness of a drug. Clinical trials provide that insight. In a legal system, an excessive focus on transparency-related information-gathering and assessment may overwhelm courts, legal practitioners, and law enforcement agencies. Meanwhile, other factors, such as measurement of effectiveness or operator competence, coupled with information on training data, may often suffice to ensure that there is a well-informed basis for trusting A/IS in a given circumstance.[129]

Given these competing considerations, arriving at a balance that is optimal for the functioning of a legal system and that has legitimacy in the eyes of the public will require an inclusive dialogue, bringing together the perspectives of those with an immediate stake in the proper functioning of a given technology, including those engaged in the design, development, procurement, deployment, operation, and validation of effectiveness of the technology, as well as those directly affected by the results of the technology; the perspectives of communities that may be indirectly impacted by the technology; and the perspectives of those with specialized expertise in ethics, government, and the law, such as jurists, regulators, and scholars. How the competing considerations should be balanced will also vary from one circumstance to another. Rather than aiming for universal transparency standards that would be applicable to all uses of A/IS within a legal system, transparency standards should allow for circumstance-dependent flexibility, in the context of the four constitutive components of trust discussed in this section.

Law

Additional note 1. The goals of transparency, e.g., answering a question as to why A/IS reached a given decision, may, in some cases, be better served by modes of explanation that do not involve examining an algorithm's terms or opening the "black box". A counterfactual explanation taking the form of, for example, "You were denied a loan because your annual income was £30,000; if your income had been £45,000, you would have been offered a loan," may provide more insight sooner than the disclosure of an algorithm.[130]

Additional note 2. The transparency principle intersects with other principles focused on fostering trust. More specifically, we note the following:

- **Transparency and effectiveness.** Information about the measurement of effectiveness can foster trust only if it is disclosed, i.e., only if there is transparency pertaining to the procedures and results of a measurement exercise.
- **Transparency and competence.** Transparency is essential in ensuring that the competencies required by the human operators of A/IS are known and met. At the same time, questions addressed by transparency extend beyond competence, while the questions addressed by competence extend beyond those answered by transparency.
- **Transparency and accountability.** Transparency is essential in determining accountability, but transparency serves purposes beyond accountability, while accountability seeks to answer questions not addressed directly by transparency.

Illustration—Transparency

In 2004, the city of Memphis, Tennessee, was experiencing an increase in crime rates that exceeded the national average. In response, in 2005, the city piloted a predictive policing program known as Blue CRUSH (Crime Reduction Utilizing Statistical History).[131] Blue CRUSH, developed in conjunction with the University of Memphis,[132] utilizes IBM's SPSS predictive analytics software to identify "hot spots": locations and times in which a given type of crime has a greater than average likelihood of occurring. The system generates its results through the analysis of a range of both historical data (type of crime, location, time of day, day of week, characteristics of victim, etc.) and live data provided by units on patrol. Equipped with the predictive crime map generated by the system, the Memphis Police Department can allocate resources dynamically to preempt or interrupt the target criminal activity. The precise response the department takes will vary with circumstance: deployment of a visible patrol car, deployment of an unmarked observer car, increasing vehicle stops in the area, undercover infiltration of the location, and so on.

The pilot program of Blue CRUSH focused on gang-related gun violence, which had been on the rise in Memphis prior to the pilot. The program showed an improvement, relative to incumbent methods, in the interdiction of such violence. Based on the success of the pilot, the scope of program was expanded, in 2007, for use throughout the city. By 2013, the policing efforts enabled by Blue CRUSH had helped to reduce overall crime in the city by over 30% and violent crime by 20%.[133] The program

Law

also enabled a dramatic increase in the rate at which crimes were solved: for cases handled by the department's Felony Assault Unit, the percentage of cases solved increased from 16% to nearly 70%.[134] And the program was cost effective: an analysis by Nucleus Research found that the program, when compared to the resources required to achieve the same results by traditional means, realized an annual benefit of approximately $7.2 million at a cost of just under $400,000.[135]

The story of the deployment of Blue CRUSH in the metropolitan Memphis area is not just about the technology; it is equally about the police personnel utilizing the technology and about the communities in which the technology was deployed. As noted by former Memphis Police Department Director Larry Godwin: "You can have all the technology in the world but you've got to have leadership, you've got to have accountability, you've got to have boots on the streets for it to succeed."[136] Crucial to the program's success was public support. Blue CRUSH represents a variety of predictive policing technology that limits itself to identifying the "where", the "when", and the "what" of criminal activity; it does not attempt to identify the "who", and therefore avoids a number of the privacy questions raised by technologies that do attempt to identify individual perpetrators. The technology will still, however, prompt responses by the police that could include more intrusive police activity in identified hot spots. The public must be willing to accept that activity, and that acceptance is won by transparency. To that end, Godwin and Janikowski held more than 200 community and neighborhood watch meetings to inform the public about the technology and how it would be used in policing their communities.[137] Without that level of transparency, it is doubtful that Blue CRUSH would have had the public support needed for its successful deployment.

Holding community meetings is an important step in building trust in a predictive policing program. As such programs become more widely implemented, however, and become more widely studied, trust may require more than town-hall meetings. Research into the programs has raised serious concerns about the ways in which they are implemented and their potential for perpetuating or even exacerbating historical bias.[138] Addressing these concerns will require more sophisticated and intrusive oversight than can be realized through community meetings.

Included among the questions that must be addressed are the following.

- In identifying hot spots, does the program rely primarily on arrest rates, which reflect (potentially biased) police activity, or does it rely on actual crime rates?
- What are the specific criteria for identifying a hot spot and are those criteria free of bias?[139]
- How accessible are the input data used to identify hot spots? Are they open to analysis by an independent expert?
- What mechanisms for oversight, review, and remediation of the program have been put in place? Such oversight should have access to the data used to train the system, the models used to identify hot spots, tests of the

Law

effectiveness of the system, and steps taken to remediate errors (such as bias) when they are uncovered.

As the public becomes more aware of the potential negative impact[140] of predictive policing programs, law enforcement agencies hoping to build trust in such programs will have to put in place transparency mechanisms that go beyond town-hall meetings and that enable a sophisticated response to such questions.

Recommendations

1. Governments and professional associations should facilitate dialogue among stakeholders—those engaged in the design, development, procurement, deployment, operation, and validation of effectiveness of the technology; those who may be immediately affected by the results of the technology; those who may be indirectly affected by the results of the technology, including the general public; and those with specialized expertise in ethics, politics, and the law—on the question of achieving a balance between transparency and other priorities, e.g., security, privacy, appropriate property rights, efficient and uniform response by the legal system, and more. In developing frameworks for achieving such balance, policymakers and professional associations should make allowance for circumstantial variation in how competing interests may be reconciled.
2. Policymakers developing frameworks for realizing transparency in A/IS applied to legal tasks should require that any frameworks they develop are sensitive both to the distinctions among the types of information that might be disclosed and to the distinctions among categories of individuals who may seek information about the design, operation, and results of a given system.
3. Policymakers developing frameworks for realizing transparency in A/IS to be adopted in a legal system should consider the role of appropriate protection for intellectual property, but should not allow those concerns to be used as a shield to prevent duly limited disclosure of information needed to ascertain whether A/IS meet acceptable standards of effectiveness, fairness, and safety. In developing such frameworks, policymakers should make allowance that the level of disclosure warranted will be, to some extent, dependent on what is at stake in a given circumstance.
4. Policymakers developing frameworks for realizing transparency in A/IS to be adopted in a legal system should consider the option of creating a role for a specially designated "public interest steward", or "trusted third party", who would be given access to sensitive information not accessible to others. Such a public interest steward would be charged with assessing the information to answer the public interest questions at hand but would be under obligation not to disclose the specifics of the information accessed in arriving at those answers.
5. Designers of A/IS should design their systems with a view to meeting transparency requirements, i.e., so as to enable some

categories of information about the system and its performance to be disclosed while enabling other categories, such as intellectual property, to be protected.

6. When negotiating contracts for the provision of A/IS products and services for use in the legal system, providers and buyers of A/IS should include contractual terms specifying what categories of information will be accessible to what categories of individuals who may seek information about the design, operation, and results of the A/IS.
7. In developing frameworks for realizing transparency in A/IS to be adopted in a legal system, policymakers should recognize that the information provided by other types of inquiries, e.g., examination of evidence of effectiveness or of operator competence, may in certain circumstances provide a more efficient means to informed trust in the effectiveness, fairness, and safety of the A/IS in question.
8. Governments should, where appropriate, work together with A/IS developers, as well as other stakeholders in the effective functioning of the legal system, to facilitate the creation of error-sharing mechanisms to enable the more effective identification, isolation, and correction of flaws in broadly deployed A/IS in their legal systems, such as a systematic facial recognition error in policing applications or in risk assessment algorithms. In developing such mechanisms, the question of precisely what information gets shared with precisely which groups may vary from application to application. All government efforts in this regard should be transparent and open to public scrutiny.
9. Governments should provide whistleblower protections to individuals who volunteer to offer information in situations where A/IS are not designed as claimed or operated as intended, or when their results are not interpreted correctly. For example, if a law enforcement agency is using facial recognition technology for a purpose that is illegal or unethical, or in a manner other than that in which it is intended to be used, an individual reporting that misuse should be given protection against reprisal. All government efforts in this regard should be transparent and open to public scrutiny.
10. Recommendation 1 under Issue 2, with respect to transparency.
11. Recommendation 2 under Issue 2, with respect to transparency.

Further Resources

- J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, "Accountable Algorithms," University of Pennsylvania Law Review, vol. 165, Feb. 2017.
- J. A. Kroll, "The fallacy of inscrutability," Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences, vol. 376, no. 2133, Oct. 2018.
- W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood, "Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations," The RAND Corporation, 2013.
- A. D. Selbst and S. Barocas, "The Intuitive Appeal of Explainable Machines," Fordham Law Review, vol. 87, no. 3, 2018.

Law

- S. Wachter, 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, vol. 7, no. 2, pp. 76-99, June 2017.
- S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," Harvard Journal of Law & Technology, vol. 31, no. 2, 2018.
- R. Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System," Stanford Law Review, vol. 70, no. 5, pp. 1342-1429, 2017.

Section 2: 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. Already, some lawmakers are wrestling with the issue of what status to apply to A/IS. Legal "personhood"—applied to humans and certain types of human organizations—is one possible option for framing such legal treatment, but granting that status to A/IS applications raises issues in multiple domains of human interaction.

Issue

What type of legal status (or other legal analytical framework) is appropriate for A/IS given (i) the legal issues raised by deployment of such technologies, and (ii) the desire to maximize the benefits of A/IS and minimize negative externalities?

Background

The convergence of A/IS and robotics technologies has led to the development of systems and devices resembling 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 resembling 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, often with an eye toward according A/IS a legal status beyond that of mere property. These entities have posited questions such as whether the law should treat such systems as legal persons.[141]

While legal personhood is a multifaceted concept, the essential feature of "full" legal personhood is the ability to participate autonomously within the legal system by having the right to sue and the capacity to be sued in court.[142] This allows legal "persons" to enter legally binding agreements, take independent action to enforce their own rights, and be held responsible for violations of the rights of others.

Conferring such status on A/IS 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 themselves human. Unlike these familiar legal entities, however, A/IS are not composed of—or necessarily controlled by—human beings. Recognizing A/IS as independent legal entities could therefore lead to abuses of that status, possibly by A/IS

Law

and certainly by the humans and legal entities who create or operate them, just as human shareholders and agents have abused the corporate form.[143] A/IS personhood is a significant departure from the legal traditions of both common law and civil law.[144]

Current legal frameworks provide a number of categories of legal status, other than full legal personhood, that could be used as analogues for the legal treatment of A/IS and how to allocate legal responsibility for harm caused by A/IS. At one extreme, legal systems could treat A/IS as mere products, tools, or other form of personal or intellectual property, and therefore subject to the applicable regimes of property law. Such treatment would have the benefit of simplifying allocation of responsibility for harm. It would, however, not account for the fact that A/IS, unlike other forms of property, may be capable of making legally significant decisions autonomously. In addition, if A/IS are to be treated as a form of property, governments and courts would have to establish rules regarding ownership, possession, and use by third parties. Other legal analogues may include the treatment of pets, livestock, wild animals, children, prisoners, and the legal principles of agency, guardianship, and powers of attorney.[145] Or perhaps A/IS are something entirely without precedent, raising the question of whether one or more types of A/IS might be assigned a hybrid, intermediate, or 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. One of the problems of exploiting the existing status of legal personhood is that international treaties may bind multiple countries to follow the lead of a single legislature, as in the EU, making it impossible for a single country to experiment with the legal and economic consequences of such a strategy.

Recognizing A/IS as independent legal persons would limit or eliminate some human responsibility for subsequent decisions made by such A/IS. For example, under a theory of intervening causation, a hammer manufacturer is not held responsible when a burglar uses a hammer to break the window of a house. However, if similar "relief" from responsibility was available to the designers, developers, and users of A/IS, it will potentially reduce their incentives to ensure the safety of A/IS they design and use. 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 legal status of A/IS thus intertwines with broader legal questions regarding how to ensure

Law

accountability and assign and allocate liability when A/IS cause harm. The question of legal personhood for A/IS, in particular, also interacts with broader ethical and practical questions on the extent to which A/IS should be treated as moral agents independent from their human designers and operators, whether recognition of A/IS personhood would enhance or detract from the purposes for which humans created the A/IS in the first place, and whether A/IS personhood facilitates of debilitates the widespread benefits of A/IS.

Some assert that because A/IS are at a very early stage of development, it is premature to choose a particular legal status or presumption in the many forms and settings in which those systems are and will be deployed. However, thoughtfully establishing a legal status early in the development could also provide crucial guidance to researchers, programmers, and developers. This uncertainty about legal status, coupled with the fact that multiple legal jurisdictions are already deploying A/IS—and each of them, as a sovereign entity, can regulate A/IS as it sees fit—suggests that there are multiple general frameworks that can and should be considered when assessing the legal status of A/IS.

Recommendations

1. While conferring full 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 now defined. 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.
2. In determining what legal status, including granting A/IS legal rights short of full legal personhood, to accord to A/IS, government and industry stakeholders alike should: (1) identify the types of decisions and operations that should never be delegated to A/IS; and (2) determine what rules and standards will most effectively ensure human control over those decisions.
3. Governments and courts should review various potential legal models—including agency, animal law, and the other analogues discussed in this section—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.
4. In addition, governments should scrutinize existing laws—especially those governing business organizations—for mechanisms that could allow A/IS to have legal autonomy. If ambiguities or loopholes create a legal method for recognizing A/IS personhood, the government should review and, if appropriate, amend the pertinent laws.
5. Manufacturers and operators should learn how each jurisdiction would categorize a given autonomous and/or intelligent system 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

Law

which that system could operate. In addition, manufacturers and operators should be aware of standards of performance and measurement promulgated by standards development organizations and agencies.

6. Stakeholders should be attentive to future developments that could warrant reconsideration of the legal status of A/IS. For example, if A/IS were developed that displayed self-awareness and consciousness, it may be appropriate to revisit the issue of whether they deserve a legal status on par with humans. Likewise, if legal systems underwent radical changes such that human rights and dignity no longer represented the primary guiding principle, the concept of full personhood for artificial entities may not represent the radical departure it might today. If the development of A/IS were to go in the opposite direction, and mechanisms were introduced allowing humans to control and predict the actions of A/IS easily and reliably, then the dangers of A/IS personhood would not be any greater than for well-established legal entities, such as corporations.
7. In considering whether to accord or expand legal protections, rights, and responsibilities to A/IS, governments should exercise utmost caution. Before according full legal personhood or a comparable legal status on A/IS, governments and courts should carefully consider whether doing so might limit how widely spread the benefits of A/IS are or could be, as well as whether doing so would harm human dignity and uniqueness of human identity. 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

- S. Bayern. "The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems." Stanford Technology Law Review 19, no. 1, pp. 93-112, 2015.
- S. Bayern, et al., "Company Law and Autonomous Systems: A Blueprint for Lawyers, Entrepreneurs, and Regulators." Hastings Science and Technology Law Journal, vol. 9, no. 2, pp. 135-162, 2017.
- D. Bhattacharyya. "Being, River: The Law, the Person and the Unthinkable." Humanities and Social Sciences Online, April 26, 2017.
- B. A. Garner, Black's Law Dictionary, 10th Edition, Thomas West, 2014.
- J. Bryson, et al., "Of, for, and by the people: the legal lacuna of synthetic persons," Artificial Intelligence Law 25, pp. 273-91, 2017.
- D. J. Calverley, "Android Science and Animal Rights, Does an Analogy Exist? " Connection Science 18, no. 4, pp. 403-417, 2006.
- D. J. Calverley, "Imagining a Non-Biological Machine as a Legal Person." AI & Society 22, pp. 403-417, 2008.
- R. Chatila, "Inclusion of Humanoid Robots in Human Society: Ethical Issues," in Springer Humanoid Robotics: A Reference, A. Goswami and P. Vadakkepat, Eds., Springer 2018.

Law

- European Parliament Resolution of 16 February 2017 (2015/2103(INL) with recommendations to the Commission on Civil Law Rules on Robotics, 2017.
- L. M. LoPucki, "Algorithmic Entities", 95 Washington University Law Review 887, 2018.
- J. S. Nelson, "Paper Dragon Thieves." Georgetown Law Journal 105, pp. 871-941, 2017.
- M. U. Scherer, "Of Wild Beasts and Digital Analogues: The Legal Status of Autonomous Systems." Nevada Law Journal 19, forthcoming 2018.
- M. U. Scherer, "Is Legal Personhood for AI Already Possible Under Current United States Laws?" Law and AI, May 14, 2017.
- L. B. Solum. "Legal Personhood for Artificial Intelligences." North Carolina Law Review 70, no. 4, pp. 1231–1287, 1992.
- J. F. Weaver. Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws. Santa Barbara, CA: Praeger, 2013.
- L. Zyga. "Incident of drunk man kicking humanoid robot raises legal questions," Techxplore, October 2, 2015.

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Law Committee

- **John Casey** (Co-Chair) – Attorney-at-Law, Corporate, Wilson Sonsini Goodrich & Rosati, P.C.
- **Nicolas Economou** (Co-Chair) – Chief Executive Officer, H5; Chair, Science, Law and Society Initiative at The Future Society; Chair, Law Committee, Global Governance of AI Roundtable; Member, Council on Extended Intelligence
- **Aden Allen** – Senior Associate, Patent Litigation, Wilson Sonsini Goodrich & Rosati, P.C.
- **Miles Brundage** – Research Scientist (Policy), OpenAI; Research Associate, Future of Humanity Institute, University of Oxford; PhD candidate, Human and Social Dimensions of Science and Technology, Arizona State University
- **Thomas Burri** – Assistant Professor of International Law and European Law, University of St. Gallen (HSG), Switzerland
- **Ryan Calo** – Assistant Professor of Law, the School of Law at the University of Washington
- **Clemens Canel** – Referendar (Trainee Lawyer) at Hanseatisches Oberlandesgericht, graduate of the University of Texas School of Law and Bucerius Law School
- **Chandramauli Chaudhuri** – Senior Data Scientist; Fractal Analytics
- **Danielle Keats Citron** – Lois K. Macht Research Professor & Professor of Law, University of Maryland Carey School of Law
- **Fernando Delgado** – PhD Student, Information Science, Cornell University.
- **Deven Desai** – Associate Professor of Law and Ethics, Georgia Institute of Technology, Scheller College of Business
- **Julien Durand** – International Technology Lawyer; Executive Director Compliance & Ethics, Amgen Biotechnology
- **Todd Elmer**, JD – Member of the Board of Directors, National Science and Technology Medals Foundation
- **Kay Firth-Butterfield** – Project Head, AI and Machine Learning at the World Economic Forum. Founding Advocate of AI-Global; Senior Fellow and Distinguished Scholar, Robert S. Strauss Center for International Security and Law, University of Texas, Austin; Co-Founder, Consortium for Law and Ethics of Artificial Intelligence and Robotics, University of Texas, Austin; Partner, Cognitive Finance Group, London, U.K.
- **Tom D. Grant** – Fellow, Wolfson College; Senior Associate of the Lauterpacht Centre for International Law, University of Cambridge, U.K.

Law

- **Cordel Green** – Attorney-at-Law; Executive Director, Broadcasting Commission—Jamaica
- **Maura R. Grossman** – Research Professor, David R. Cheriton School of Computer Science, University of Waterloo; Adjunct Professor, Osgoode Hall Law School, York University
- **Bruce Hedin** – Principal Scientist, H5
- **Daniel Hinkle** – Senior State Affairs Counsel for the American Association for Justice
- **Derek Jinks** – Marrs McLean Professor in Law, University of Texas Law School; Director, Consortium on Law and Ethics of Artificial Intelligence and Robotics (CLEAR), Robert S. Strauss Center for International Security and Law, University of Texas.
- **Nicolas Jupillat** – Adjunct Professor, University of Detroit Mercy School of Law
- **Marwan Kawadri** – Analyst, Founders Intelligence; Research Associate, The Future Society.
- **Mauricio K. Kimura** – Lawyer; PhD student, Faculty of Law, University of Waikato, New Zealand; LLM from George Washington University, Washington DC, USA; Bachelor of Laws from Sao Bernardo do Campo School of Law, Brazil
- **Irene Kitsara** – Lawyer; IP Information Officer, Access to Information and Knowledge Division, World Intellectual Property Organization, Switzerland
- **Timothy Lau**, **J.D., Sc.D.** – Research Associate, Federal Judicial Center
- **Mark Lyon** – Attorney-at-Law, Chair, Artificial Intelligence and Autonomous Systems Practice Group at Gibson, Dunn & Crutcher LLP
- **Gary Marchant** – Regents' Professor of Law, Lincoln Professor of Emerging Technologies, Law and Ethics, Arizona State University
- **Nicolas Miailhe** – Co-Founder & President, The Future Society; Member, AI Expert Group at the OECD; Member, Global Council on Extended Intelligence; Senior Visiting Research Fellow, Program on Science Technology and Society at Harvard Kennedy School. Lecturer, Paris School of International Affairs (Sciences Po); Visiting Professor, IE School of Global and Public Affairs
- **Paul Moseley** – Master's student, Electrical Engineering, Southern Methodist University; graduate of the University of Texas School of Law
- **Florian Ostmann** – Policy Fellow, The Alan Turing Institute
- **Pedro Pavón** – Assistant General Counsel, Global Data Protection, Honeywell
- **Josephine Png** – AI Policy Researcher and Deputy Project Manager, The Future Society; budding barrister; and BA Chinese and Law, School of Oriental and African Studies
- **Matthew Scherer** – Attorney at Littler Mendelson, P.C., and legal scholar based in Portland, Oregon, USA; Editor, LawAndAI.com
- **Bardo Schettini Gherardini** – Independent Legal Advisor on standardization, AI and robotics

Law

- **Jason Tashea** – Founder, Justice Codes and adjunct law professor at Georgetown Law Center
- **Yan Tougas** – Global Ethics & Compliance Officer, United Technologies Corporation; Adjunct Professor, Law & Ethics, University of Connecticut School of Business; Fellow, Ethics & Compliance Initiative; Kallman Executive Fellow, Bentley University Hoffman Center for Business Ethics
- **Sandra Wachter** – Lawyer and Research Fellow in Data Ethics, AI and Robotics, Oxford Internet Institute, University of Oxford
- **Axel Walz** – Lawyer; Senior Research Fellow at the Max Planck Institute for Innovation and Competition, Germany. (Member until October 31, 2018)
- **John Frank Weaver** – Lawyer, McLane Middleton, P.A; Columnist for and Member of Board of Editors of *Journal of Robotics, Artificial Intelligence & Law*; Contributing Writer for *Slate*; Author, *Robots Are People Too*
- **Julius Weitzdörfer** – Affiliated Lecturer, Faculty of Law, University of Cambridge; Research Associate, Centre for the Study of Existential Risk, University of Cambridge
- **Yueh-Hsuan Weng** – Assistant Professor, Frontier Research Institute for Interdisciplinary Sciences (FRIS), Tohoku University; Fellow, Transatlantic Technology Law Forum (TTLF), Stanford Law School
- **Andrew Woods** – Associate Professor of Law, University of Arizona

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

Law

The Law Committee of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems would like to thank the following individuals for taking the time to offer valuable feedback and suggestions on Section 1 of the Law Chapter, "Norms for the Trustworthy Adoption of A/IS in Legal Systems". Each of these contributors offered comments in an individual capacity, not in the name of the organization for which they work. The final version of the Section does not necessarily incorporate all comments or reflect the views of each contributor.

- **Rediet Abebe**, PhD Candidate, Department of Computer Science, Cornell University; cofounder, Mechanism Design for Social Good; cofounder, Black in AI.
- **Ifeoma Ajunwa**, Assistant Professor, Labor & Employment Law, Cornell Industrial and Labor Relations School; faculty Associate at Harvard Law, Berkman Klein Center.
- **Jason R. Baron**, of counsel, Drinker Biddle; co-chair, Information Governance Initiative; former Director of Litigation, United States National Archives and Records Administration.
- **Irakli Beridze**, Head, Centre for Artificial Intelligence and Robotics, United Nations (UNICRI).
- **Juan Carlos Botero**, Law Professor, Pontificia Universidad Javeriana, Bogota; former Executive Director, World Justice Project.
- **Anne Carblanc**, Principal Administrator, Information, Communications and Consumer Policy (ICCP) Division, Directorate for Science, Technology and Industry, OECD; former criminal investigations judge (juge d'instruction), Tribunal of Paris.
- **Gallia Daor**, Policy Analyst, OECD.
- **Lydia de la Torre**, Privacy Law Fellow, Santa Clara University.
- **Isabela Ferrari**, Federal Judge, Federal Court, Rio de Janeiro, Brazil.
- **Albert Fox Cahn**, Founder and Executive Director, Surveillance Technology Oversight Project; former Legal Director, CAIR-NY.
- **Paul W. Grimm**, United States District Judge, United States District Court for the District of Maryland.
- **Gillian Hadfield**, Professor of Law and Professor of Strategic Management, University of Toronto; Member, World Economic Forum Future Council for Agile Governance.
- **Sheila Jasanoff**, Pforzheimer Professor of Science and Technology Studies, Harvard Kennedy School of Government.
- **Baroness Beeban Kidron**, OBE, Member, United Kingdom House of Lords.
- **Eva Kaili**, Member, European Parliament; Chair, European Parliament Science and Technology Options Assessment body (STOA).
- **Mantalena Kaili**, cofounder, European Law Observatory on New Technologies.
- **Jon Kleinberg**, Tisch University Professor, Departments of Computer Science and Information Science, Cornell University; member of the National Academy of Sciences, the National Academy of Engineering, and the American Academy of Arts and Sciences.
- **Shuang Lu Frost**, Teaching Fellow, PhD candidate, Department of Anthropology, Harvard University.

Law

- **Arthur R. Miller CBE**, University Professor, New York University; former Bruce Bromley Professor of Law, Harvard Law School.
- **Manuel Muñiz**, Dean and Rafael del Pino Professor of Practice of Global Leadership, IE School of Global and Public Affairs, Madrid; Senior Associate, Belfer Center, Harvard University.
- **Erik Navarro Wolkart**, Federal Judge, Federal Court, Rio de Janeiro, Brazil.
- **Aileen Nielsen**, chair, Science and Law Committee, New York City Bar Association.
- **Michael Philips**, Assistant General Counsel, Microsoft.
- **Dinah PoKempner**, General Counsel, Human Rights Watch.
- **Irina Raicu**, Director, Internet Ethics Program, Markkula Center for Applied Ethics, Santa Clara University.
- **David Robinson**, Visiting Scientist, AI Policy and Practice Initiative, Cornell University; Adjunct Professor of Law, Georgetown University Law Center; Managing Director (on leave), Upturn.
- **Alanna Rutherford**, Vice President, Global Litigation & Competition, Visa.
- **George Socha, Esq.**, Consulting Managing Director, BDO USA; co-founder, Electronic Discovery Reference Model (EDRM) and Information Governance Reference Model (IGRM).
- **Lee Tiedrich**, Partner, IP/Technology Transactions, and Co-Chair, Artificial Intelligence Initiative, Covington & Burling LLP.
- **Darrell M. West**, VP, Governance Studies, Director, Center for Technology Innovation, Douglas Dillon Chair in Governance Studies, Brookings Institution.
- **Bendert Zevenbergen**, Research Fellow, Center for Information Technology Policy, Princeton University; Researcher, Oxford Internet Institute.
- **Jiyu Zhang**, Associate Professor and Executive Director of the Law and Technology Institute, Renmin University of China School of Law.
- **Peter Zimroth**, Director, New York University Center on Civil Justice; retired partner, Arnold & Porter; former Assistant US Attorney, Southern District of New York.

Endnotes

[1] See S. Jasanoff, "Governing Innovation: The Social Contract and the Democratic Imagination," Seminar, vol. 597, pp. 16-25, May 2009.

[2] As articulated in *EAD* General Principles 1 (Human Rights), 2 (Well-Being), and 3 (Data Agency). See also *EAD* Chapter, "Classical Ethics in A/IS," In applying A/IS in pursuit of these goals, tradeoffs are inevitable. Some applications of predictive policing, for example, may reduce crime, and so enhance well-being, but may do so at the cost of impinging on a right to privacy or weakening protections against unwarranted search and seizure. How these tradeoffs are negotiated may vary with cultural and legal traditions.

[3] Risks and benefits, and their perception, are neither always well-defined at the outset nor static over time. Social expectations and even ideas of lawfulness constantly evolve. For example, if younger generations, accustomed to the use of social networking technologies, have lower expectations of privacy than older generations, should this be deemed to be a benefit to society, a risk, or neither?

[4] Regarding the nature of the guidance provided in this section: Artificial intelligence, like many other domains relied on by the legal realm (e.g., medical and accounting forensics, ballistics, or economic analysis), is a scientific discipline distinct from the law. Its effective and safe design and operation have underpinnings in academic and professional competencies in computer science, linguistics, data science, statistics, and related technical fields. Lawyers, judges, and law enforcement officers increasingly draw on these fields, directly or indirectly, as A/IS are progressively adopted in the legal system. This document does not seek to offer legal advice to lawyers, courts, or law enforcement agencies on how to practice their professions or enforce the law in their jurisdictions around the globe. Instead, it seeks to help ensure that A/IS and their operators in a given legal system can be trusted by lawyers, courts, and law enforcement agencies, and civil society at large, to perform effectively and safely. Such effective and safe operation of A/IS holds the potential of producing substantial benefits for the legal system, while protecting all of its participants from the ethical, professional, and business risks, or personal jeopardy, that may result from the intentional, unintentional, uninformed, or incompetent procurement and operation of artificial intelligence.

[5] See Rensselaer Polytechnic Institute, "A Conversation with Chief Justice John G. Roberts, Jr.," April 11, 2017. YouTube video, 40:12. April 12, 2017. [Online]. Available: https://www.youtube.com/watch?v=TuZEKlRgDEg.

[6] "Uninformed avoidance of adoption" can be one of two types: (a) avoidance of adoption when the information needed to enable sound decisions is available but is not taken into

consideration, and (b) avoidance of adoption when the information needed to enable sound decisions is simply not available. Unlike the former type of avoidance, the latter type is a prudent and well-reasoned avoidance of adoption and, pending better information, is the course recommended by a number experts and nonexperts.

[7] For purposes of this chapter, we have made the deliberate choice to focus on these four principles without taking a prior position on where the deployment of A/IS may or may not be acceptable in legal systems. Where these principles cannot be adequately operationalized, it would follow that the deployment of A/IS in a legal system cannot be trusted. Where A/IS can be evidenced to meet desired thresholds for each duly operationalized principle, it would follow that their deployment can be trusted. Such information is intended to facilitate, not preempt, the indispensable public policy dialogue on the extent to which A/IS should be relied upon to meet the specific needs of the legal systems of societies around the world.

[8] It is beyond the scope of this chapter to discuss the process through which such adherence may become institutionalized in the complex legal, technological, political, and cultural dynamics in which sociotechnical innovation occurs. It is worth noting, however, that this process typically involves four steps. First, a wide range of market and culture-driven practices emerge. Second, a set of best practices arises, reflecting a group's willingness to adopt certain rules. Third, some of these best practices are formulated into standards, which enable enforcement (through private contracts, professional codes of practice, or legislation). Finally, those enforceable standards render the performance of some activities sufficiently reliable and predictable to enable trustworthy operation at the scale of society. Where these elements (rulemaking, enforcement, scalable operation) are present, new institutions are born.

[9] For a discussion of the definition of A/IS, see the Terminology Update in the Executive Summary of EAD. The principles outlined in this section as constitutive of "informed trust" do not depend on a precise, consensus definition of A/IS and are, in fact, designed to be enable successful operationalization under a broad range of definitions.

[10] Such as Gross Domestic Product (GDP), Gross National Income (GNI) per capita, the WEF Global Competitiveness Index, and others.

[11] Such as life expectancy, infant mortality rate, and literacy rate, as well as composite indices such as the Human Development Index, the Inequality-Adjusted Human Development Index, the OECD Framework for Measuring Well-being and Progress, and others. For more on measures of well-being, see the *EAD* chapter on "Well-being".

[12] See United Nation General Assembly, Universal Declaration of Human Rights, Dec. 10, 1948, available: http://www.un.org/en/universal-declaration-human-rights/index.html; see also United Nations Office of the High Commissioner: Human Rights, The Vienna Declaration and Programme of Action, June 25, 1993, available: https://www.ohchr.org/en/professionalinterest/pages/vienna.aspx.

Law

[13] See UNICEF, Convention on the Rights of the Child, Nov. 4, 2014, available: https://www.unicef.org/crc/index_30160.html.

[14] See United Nations Security Council, "The Rule of Law and Transitional Justice in Conflict and Post-conflict Societies: Report of the Secretary General," *Report S/2004/616* (2004).

[15] See The World Economic Forum, *The Global Competitiveness Report: 2018*, ed. K. Schwab (2018), pp. 12ff.

[16] See A. Brunetti, G. Kisunko, and B. Weder, "Credibility of Rules and Economic Growth: Evidence from a Worldwide Survey of the Private Sector," The World Bank Economic Review, vol. 12, no. 3, pp. 353–384, 1998. Available: https://doi.org/10.1093/wber/12.3.353; see also World Bank, World Development Report 2017: Governance and the Law, Jan. 2017. Available: doi.org/10.1596/978-1-4648-0950-7.

[17] The question of intellectual property law in an era of rapidly advancing technology (both A/IS and other technologies) is a complex and often contentious one involving legal, economic, and ethical considerations. We have not yet studied the question in sufficient depth to reach a consensus on the issues raised. We may examine the issues in depth in a future version of *EAD*. For a forum in which such issues are discussed, see the Berkeley-Stanford Advanced Patent Law Institute. See also The World Economic Forum, "Artificial Intelligence Collides with Patent Law." April 2018. Available: http://www3.weforum.org/docs/WEF_48540_WP_End_of_Innovation_Protecting_Patent_Law.pdf.

[18] A component of human dignity is privacy, and a component of privacy is protection and control of one's data; in this regard, frameworks such as the EU's General Data Protection Regulation (GDPR) and the Council of Europe's "Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data" have a role to play in setting standards for how legal systems can protect data privacy. See also *EAD* General Principle 3 (Data Agency).

[19] Frameworks such as the Universal Declaration of Human Rights and the Vienna Declaration and Programme of Action (VDPA) have a role to play in articulating human-rights standards to which legal systems should adhere. See also *EAD* General Principle 1 (Human Rights).

[20] For more on the importance of measures of well-being beyond GDP, see *EAD* General Principle 2 (Well-being).

[21] For a conceptual framework enabling the country-by-country assessment of the Rule of Law, see World Justice Project, *Rule of Law Index*. 2018. url: https://worldjusticeproject.org/sites/default/files/documents/WJP-ROLI-2018-June-Online-Edition_0.pdf.

[22] *See* D. Kennedy, "The 'Rule of Law,' Political Choices and Development Common Sense," in The New Law and Economic Development: A Critical Appraisal, D. M. Trubek and A. Santos, Ed. Cambridge: Cambridge University Press, 2006, pp. 156-157; see also A. Sen, *Development as Freedom*. New York: Alfred A. Knopf, 1999.

Law

[23] *See* Kennedy (2006): pp. 168-169. "The idea that building 'the rule of law' might *itself* be a development strategy encourages the hope that choosing law *in general* could substitute for all the perplexing political and economic choices that have been at the center of development policy making for half a century. The politics of allocation is submerged. Although a legal regime offers an arena to contest those choices, it cannot substitute for them."

[24] *Fairness* (as well as *bias*) can be defined in more than one way. For purposes of this chapter, a commitment is not made to any one definition—and indeed, it may not be either desirable or feasible to arrive at a single definition that would be applied in all circumstances. The trust principles proposed in the chapter (Effectiveness, Competence, Accountability, and Transparency) are defined such that they will provide information that will allow the testing of an application of A/IS against any fairness criteria.

[25] The confidentiality of jury deliberations, certain sensitive cases, and personal data are some of the considerations that influence the extent of appropriate public examination and oversight mechanisms.

[26] The avoidance of negative consequences is important to note in relation to effectiveness. The law can be used for malevolent or intensely disputed purposes (for example, the quashing of dissent or mass incarceration). The instruments of the law, including A/IS, can render the advancement of such purposes more effective to the detriment of democratic values, human rights, and human well-being.

[27] Studies conducted by the US National Institute of Standards and Technology (NIST) between 2006 and 2011, known as the US NIST Text REtrieval Conference (TREC) Legal Track, suggest that some A/IS-enabled processes, if operated by trained experts in the relevant scientific fields, can be more effective (or accurate) than human attorneys in correctly identifying case-relevant information in large data sets. NIST has a long-standing reputation for cultivating trust in technology by participating in the development of standards and metrics that strengthen measurement science and make technology more secure, usable, interoperable, and reliable. This work is critical in the A/IS space to ensure public trust of rapidly evolving technologies so that we can benefit from all that this field has to promise.

[28] In describing the potential A/IS have for aiding in the auditing of decisions made in the civil and criminal justice systems, we are envisioning them acting as aids to a competent human auditor (see Issue 4) in the context of internal or judicial review.

[29] Of course, the use of A/IS in improving the effectiveness of law enforcement may raise concerns about other aspects of well-being, such as privacy and the rise of the surveillance state, cf. Minority Report (2002). If A/IS are to be used for law enforcement, steps must be taken to ensure that they are used, and that citizens trust that they will be used, in ways that are conducive to ethical law enforcement and individual well-being (see Issue 2).

Law

[30] A/IS may also provide assistance in carrying out legal tasks associated with larger transactions, such as evaluating contracts for risk in connection with a M&A transaction or reporting exposure to regulators.

[31] The recommendations provided in this chapter (both under this issue and under the other issues discussed in the chapter) are intended to give general guidance as to how those with a stake in the just and effective operation of a legal system can develop norms for the trustworthy adoption of A/IS in the legal system. The specific ways in which the recommendations are operationalized will vary from society to society and from jurisdiction to jurisdiction.

[32] See "Global Governance of AI Roundtable: Summary Report 2018," World Government Summit, 2018: p. 32. Available: https://www.worldgovernmentsummit.org/api/publications/document?id=ff6c88c5-e97c-6578-b2f8-ff0000a7ddb6. (The February 2018 Dubai Global Governance of AI Roundtable brought together ninety leading thinkers on AI governance.)

[33] *See State v Loomis*, 881 N.W.2d 749 (Wis. 2016), *cert. denied* (2017); see also "Criminal Law—Sentencing Guidelines—Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing—State v. Loomis, 881 N.W.2d 749 (Wis. 2016)," Harvard Law Review, vol. 130, no. 5, pp. 1535-1536, 2017. Available: http://harvardlawreview.org/wp-content/uploads/2017/03/1530-1537_online.pdf; see also K. Freeman, "Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis," North Carolina Journal of Law and Technology, vol. 18, no. 5, pp. 75-76, 2016. Available: https://scholarship.law.unc.edu/ncjolt/vol18/iss5/3/.

[34] An example of an initiative that seeks to bridge the gap between technical and legal expertise is the Artificial Intelligence Legal Challenge, held at Ryerson University and sponsored by Canada's Ministry of the Attorney General: http://www.legalinnovationzone.ca/press_release/ryersons-legal-innovation-zone-announces-winners-of-ai-legal-challenge/.

[35] And, in addressing the challenges, consideration must be given to existing modes of proposing and approving innovation in the legal system. Trust in A/IS will be undermined if they are viewed as not having been vetted via established processes.

[36] For an overview of risk and risk management, see Working Party on Security and Privacy in the Digital Economy, Background Report for Ministerial Panel 3.2, Directorate for Science, Technology and Innovation, Committee on Digital Economy Policy, Managing Digital Security and Privacy Risk, OECD, June 1, 2016; see p. 5.

[37] It is worth emphasizing the "informed" qualifier we attach to trust here. Far from advocating for a "blind trust" in A/IS, we argue that A/IS should be adopted only when we have sound evidence of their effectiveness, when we can be confident of the competence of their operators, when we have assurances that these systems allow for the attribution of responsibility for outcomes (both positive and negative), and when we have clear views into their operation. Without those conditions, we would argue that A/IS *should not be adopted* in the legal system.

Law

[38] The importance of testing the effectiveness of advanced technologies, including A/IS, in the legal system (and beyond) is not new: it was highlighted by Judge Paul W. Grimm in an important early ruling on legal fact-finding, *Victor Stanley v. Creative Pipe, Inc.*, 250 F.R.D. 251, 257 (D. Md. 2008), followed, among others, by the influential research and educational institute The Sedona Conference as well as the International Organization for Standardization (ISO). See *An Open Letter to Law Firms and Companies in the Legal Tech Sector*, The Sedona Conference (2009), and *Commentary on Achieving Quality in the E-Discovery Process* (2013): 7; ISO standard on electronic discovery (ISO/IEC 27050-3:2017): 19. Most recently, in the summary report of the Global Governance of AI Roundtable at the 2018 World Government Summit, Omar bin Sultan Al Olama, Minister of State for Artificial Intelligence of the UAE, highlighted the importance of "empirical information" in assessing the suitability of A/IS.

[39] In the terminology of software development, *verification* is a demonstration that a given application meets a narrowly defined requirement; *validation* is a demonstration that the application answers its real-world use case. When we speak of gathering evidence of the effectiveness of A/IS, we are speaking of validation.

[40] Standards may include compliance with defined professional competence or other ethical requirements, but also other types of standards, such as data standards. Data standards may serve as "a digital lingua franca" with the potential of both supporting broad-based technological innovation (including A/IS innovation) in a legal system and facilitating access to justice. As part of interactive technology solutions, appropriate data standards may help connect the ordinary citizen to the appropriate resources and information for his or her legal needs. For a discussion of open data standards in the context of the US court system, see D. Colarusso and E. J. Rickard, "Speaking the Same Language: Data Standards and Disruptive Technologies in the Administration of Justice," Suffolk University Law Review, vol. L387, 2017.

[41] For measurement of bias in facial recognition software, see C. Garvie, A. M. Bedoya, and J. Frankle, "The Perpetual Line-Up: Unregulated Police Face Recognition in America," Georgetown Law, Center on Privacy & Technology, Oct. 2016. Available: https://www.perpetuallineup.org/.

[42] The inclusion of such collateral effects in assessing effectiveness is an important element in overcoming the apparent "black box" or inscrutable nature of A/IS. See, for example, J. A. Kroll, "The fallacy of inscrutability," Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences, vol. 376, no. 2133, Oct. 2018. Available: doi.org/0.1098/rsta.2018.0084. The study addresses, among other questions, "how measurement of a system beyond understanding of its internals and its design can help to defeat inscrutability."

[43] The question of the salience of collateral impact will vary with the specific application of A/IS. For example, false positives in document review related to fact-finding will generally not raise acute ethical issues, but false positives

in predictive policing or sentencing will. In these latter domains, complex and sometimes unsettled issues of fairness arise, particularly when social norms of fairness change regionally and over time (sometimes rapidly). Any A/IS that was designed to replicate some notion of fairness would need to demonstrate its effectiveness, first, at replicating prevailing notions of fairness that have legitimacy in society, and second, at responding to evolutions in such notions of fairness. In the current state of A/IS, in which no system has been able to demonstrate consistent effectiveness in either of the above regards, it is essential that great discretion be exercised in considering any reliance on A/IS in domains such as sentencing and predictive policing.

[44] These exercises go by various names in the literature: *effectiveness evaluations*, *benchmarking exercises*, *validation studies*, and so on. See, for example, the definition of *validation study* in AINOW's 2018 *Algorithmic Accountability Toolkit* (https://ainowinstitute.org/aap-toolkit.pdf), p. 29. For our purposes, what matters is that the exercise be one that collects, in a scientifically sound manner, evidence of how "fit for purpose" any given A/IS are.

[45] This feature of evaluation design is important, as only tasks that accurately reflect real-world conditions and objectives (which may include the avoidance of unintended consequences, such as racial bias) will provide compelling guidance as to the suitability of an application for adoption in the real world.

[46] For TREC generally, see: https://trec.nist.gov/. For the TREC Legal Track specifically, see: https://trec-legal.umiacs.umd.edu/.

[47] When a complex system can be broken down into separate component systems, it may be appropriate to assess either the effectiveness of each component, or that of the end-to-end application as a whole (including human operators), depending on the specific question to be answered.

[48] Qualitative considerations may also help counter attempts to "game the system" (i.e., attempts to use bad-faith methods to meet a specific numerical target); see B. Hedin, D. Brassil, and A. Jones, "On the Place of Measurement in E-Discovery," in Perspectives on Predictive Coding and Other Advanced Search Methods for the Legal Practitioner, ed. J. R. Baron, R. C. Losey, and M. D. Berman. Chicago: American Bar Association, 2016, p. 415f.

[49] Even in fact-finding, accurate extraction of facts does not eliminate the need for reasoned judgment as to the significance of the facts in the context of specific circumstances and cultural considerations. Used properly, A/IS will advance the spirit of the law, not just the letter of the law.

[50] Electronic discovery is the task of searching through large collections of electronically stored information (ESI) for material relevant to civil and criminal litigation and investigations. Among applications of A/IS to legal tasks and questions, the application to legal discovery is probably the most "mature," as measured against the criteria of having been tested, assessed and approved by courts, and adopted fairly widely across various jurisdictions.

Law

[51] While there is general consensus about the importance of these metrics in gauging effectiveness in legal discovery, there is not a consensus around the precise values for those metrics that must be met for a discovery effort to be acceptable. That is a good thing, as the precise value that should be attained, and demonstrated to have been attained, in any given matter will be dependent on, and proportional to, the specific facts and circumstances of that matter.

[52] Different domains of application of A/IS to legal matters will vary not only with regard to the availability of consensus metrics of effectiveness, but also with regard to conditions that affect the challenge of measuring effectiveness: availability of data, impact of social bias, and sensitivity to privacy concerns all affect how difficult it may be to arrive at consensus protocols for gauging effectiveness. In the case of defining an effectiveness metric for A/IS used in support of sentencing decisions, one challenge is that, while it is easy to find when an individual who has been released commits a crime (or is convicted of committing a crime), it is difficult to assess when an individual who was not released would have committed a crime. For a discussion of the challenges in measuring the effectiveness of tools designed to assess flight risk, see M. T. Stevenson, "Assessing Risk Assessment in Action." Minnesota Law Review, vol. 103, 2018. Available: doi.org/10.2139/ssrn.3016088.

[53] Sound measurement may also serve as an effective antidote to the unsubstantiated claims sometimes made regarding the effectiveness of certain applications of A/IS to legal matters (e.g., flight risk assessment technologies); see Stevenson, "Assessing Risk Assessment". Unsubstantiated claims are an appropriate source of an *informed distrust* in A/IS. Such well-founded distrust can be addressed only with truly meaningful and sound measures that provide accurate information regarding the capabilities and limitations of a given system.

[54] See the discussion under "Illustration—Effectiveness" in this chapter.

[55] For more on principles for data protection, see the EAD chapter "Personal Data and Individual Agency".

[56] The importance of validation by practitioners is reflected in The European Commission's High-Level Expert Group on Artificial Intelligence Draft Ethics Guidelines for Trustworthy AI: "Testing and validation of the system should thus occur as early as possible and be iterative, ensuring the system behaves as intended throughout its entire life cycle *and especially after deployment.*" (Emphasis added.) See High-Level Expert Group on Artificial Intelligence, "DRAFT Ethics Guidelines for Trustworthy AI: Working Document for Stakeholders' Consultation," The European Commission. Brussels, Belgium: Dec. 18, 2018.

[57] That scrutiny need not extend to IP or other protected information (e.g., attorney work product). Validation methods and results are a matter of numbers and procedures for obtaining the numbers, and their disclosure would not impinge on safeguards against the disclosure of legitimately protected information.

Law

[58] A recent matter from the US legal system illustrates how a failure to disclose the results of a validation exercise can limit the exercise's ability to achieve its intended purpose. In *Winfield v. City of New York* (Opinion & Order. 15-CV-05236 [LTS] [KHP]. SDNY 2017), a party had utilized the A/IS-enabled system to conduct a review of documents for relevance to the matter being litigated. When the accuracy and completeness of the results of that review were challenged by the requesting party, the producing party disclosed that it had, in fact, conducted validation of its results. Rather than requiring that the producing party simply disclose the results of the validation to the requesting party, the judge overseeing the dispute chose to review the results herself *in camera*, without providing access to the requesting party. Although the judge then said that the evidence she was provided supported the accuracy and completeness of the review, the requesting party could not itself examine either the evidence or the methods whereby it was obtained, and so could not gain confidence in the results. That confidence comes only from examining the metrics and the procedures followed in obtaining them. Moreover, the results of a validation exercise, which are usually simple numbers that reflect sampling procedures, can be disclosed without revealing the content of any documents, any proprietary tools or methods, or any attorney work product. If the purpose of conducting a validation exercise is to gather evidence of the effectiveness of a process, in the event that the process is challenged, keeping that evidence hidden from those who would challenge the process limits the ability of the validation exercise to achieve its intended purpose.

[59] https://www.nist.gov/.

[60] TREC Legal Track (2006-2011): https://trec-legal.umiacs.umd.edu/.

[61] The statistical evidence in question here is statistical evidence of the effectiveness of A/IS applied to the task of discovery; it is not statistical evidence of facts actually at issue in litigation. Courts may have different rules for the admissibility of the two kinds of statistical evidence (and there will be jurisdictional differences on these questions).

[62] It is important to underscore that, whereas developers and operators of A/IS should be able to derive sound measurements of effectiveness, the courts should determine what level of effectiveness—what score—should be demonstrated to have been achieved, based on the facts and circumstances of a given matter. In some instances, the cost (in terms of sample sizes, resources required to review the samples, and so on) of demonstrating the achievement of a high score will be disproportionate to the stakes of a given matter. In others, for example, a major securities fraud claim that potentially affects thousands of citizens, a court might justifiably demand a demonstration of the achievement of a very high score, irrespective of cost. Demonstrations of the effectiveness of A/IS (and of their operators) are instruments in support of, not in substitution of, judicial decision-making.

[63] See, for example, B. Hedin, S. Tomlinson, J. R. Baron, and D. W. Oard, "Overview of the TREC 2009 Legal Track," in *NIST Special Publication: SP 500-278, The Eighteenth Text REtrieval Conference (TREC 2009) Proceedings* (2009).

Law

[64] See M. R. Grossman 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, vol. 17, no. 3, 2011. Available: http://jolt.richmond.edu/jolt-archive/v17i3/article11.pdf. Note that the two systems that conclusively demonstrated "better than human" performance took methodologically distinct approaches, but they shared the characteristic of having been designed, operated, and measured for accuracy by scientifically trained experts.

[65] *Da Silva Moore v. Publicis Groupe*, 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012). See also A. Peck, "Search, Forward," Legaltech News. Oct. 1, 2011. Available: https://www.law.com/legaltechnews/almID/1202516530534Search-Forward/.

[66] The fact that NIST has as important role to play in developing standards for the measurement of the safety and security of A/IS was recognized in a recent (September, 2018) report from the U.S. House of Representatives: "At minimum, a widely agreed upon standard for measuring the safety and security of AI products and applications should precede any new regulations. ... The National Institute of Standards and Technology (NIST) is situated to be a key player in developing standards." (Will Hurd and Robin Kelly, "Rise of the Machines: Artificial Intelligence and its Growing Impact on U.S. Policy," U.S. House of Representatives—Committee on Oversight and Government Reform—Subcommittee on Information Technology, September, 2018).

[67] The competence principle is intended to apply to the post design operation of A/IS. Of course, that does not mean that designers and developers of A/IS are free of responsibility for their systems' outcomes. As discussed in the background to this issue, it is incumbent on designers and developers to assess the risks associated with the operation of their systems and to specify the operator competencies needed to mitigate those risks. For more on the question of designer incompetence or negligence, see the discussion of "software malpractice" in Kroll (2018).

[68] The ISO standard on e-discovery, ISO/IEC 27050-3, does recognize the importance of expertise in applying advanced technologies in a search for documents responsive to a legal inquiry; see ISO/IEC 27050-3: *Information technology — Security techniques — Electronic discovery — Part 3: Code of practice for electronic discovery*, Geneva (2017), pp. 19-20.

[69] See, for example, ABA Model Rule 1, comment 8: "To maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology, engage in continuing study and education and comply with all continuing legal education requirements to which the lawyer is subject." Available: https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_1_competence/comment_on_rule_1_1/. See also, The State Bar of California Standing Committee on Professional Responsibility and Conduct, Formal Opinion No. 2015-193. Available:

Law

https://www.calbar.ca.gov/Portals/0/documents/ethics/Opinions/CAL%202015-193%20%5B11-0004%5D%20(06-30-15)%20-%20FINAL.pdf.

[70] In the deliberations of the Law Committee of the 2018 Global Governance of AI Roundtable, the question of the competencies needed "in order to effectively operate and measure the efficacy of AI systems in legal functions that affect the rights and liberty of citizens" was cited as one of the considerations that "appear to be most overlooked in the current public dialogue." See "Global Governance of AI Roundtable: Summary Report 2018," World Government Summit, 2018: p. 7. Available: https://www.worldgovernmentsummit.org/api/publications/document?id=ff6c88c5-e97c-6578-b2f8-ff0000a7ddb6.

[71] See A. G. Ferguson, "Policing Predictive Policing," Washington University Law Review, vol. 94, no. 5, 2017: 1109, 1172. Available: https://openscholarship.wustl.edu/law_lawreview/vol94/iss5/5/.

[72] In addition, a lack of competence in interpreting the results of a statistical exercise can (and often does) result in an incorrect conclusion (on the part of a party to a dispute or of a judge seeking to resolve a dispute). For example, in *In re: Biomet*, a judge addressing a discovery dispute interpreted the statistical data provided by the producing party as indicating that the producing party's retrieval process had left behind "a comparatively modest number" of responsive documents, when the statistical evidence showed, in fact, that a substantial number of responsive documents had been left behind. See *In re: Biomet M2a Magnum Hip Implant Prods. Liab. Litig.*No. 3:12-MD-2391 (N.D. Ind. April 18, 2013).

[73] For example, a prior violent conviction may be weighted equally, whether the violent act was a shove or a knife attack. See Human Rights Watch. "Q & A: Profile Based Risk Assessment for US Pretrial Incarceration, Release Decisions," June 1, 2018. Available: https://www.hrw.org/news/2018/06/01/q-profile-based-risk-assessment-us-pretrial-incarceration-release-decisions.

[74] Bias can be introduced in a number of ways: via the features taken into consideration by the algorithm, via the nature and composition of the training data, via the design of the validation protocol, and so on. A competent operator will be alert to and assess such potential sources of bias.

[75] Among the conditions may be, for example, that the results of the system are to be used only to provide guidance to the human decision maker (e.g., judge) and should not be taken as, in themselves, dispositive.

[76] Given that the effective functioning of a legal system is a matter of interest to the whole of society, it is important that all members of a society be provided with access to the resources needed to understand when and how A/IS are applied in support of the functioning of a legal system.

[77] Among the topics covered by such training should be the potential for "automation bias" and ways to mitigate it. See L. J. Skitka, K. Mosier, and M. D. Burdick, "Does automation

bias decision-making?" International Journal of Human-Computer Studies, vol. 51, no. 5, pp. 991-1006, 1999. Available: https://doi.org/10.1006/ijhc.1999.0252; L. J. Skitka, K. Mosier, and M. D. Burdick, "Accountability and automation bias," International Journal of Human-Computer Studies, vol. 52, no. 4, pp. 701-717, 2000. Available: https://doi.org/10.1006/ijhc.1999.0349.

[78] Some government agencies are working toward creating a more effective partnership between the skills found in technology start-ups and the skills required of legal practitioners. See Legal Innovation Zone. "Ryerson's Legal Innovation Zone Announces Winners of AI Legal Challenge," March 26, 2018. Available: http://www.legalinnovationzone.ca/press_release/ryersons-legal-innovation-zone-announces-winners-of-ai-legal-challenge/.

[79] See Amazon. "Amazon Rekognition." https://aws.amazon.com/rekognition/ (2018).

[80] See E. Dwoskin, "Amazon is selling facial recognition to law enforcement—for a fistful of dollars." *Washington Post*, May 22, 2018. Available: https://www.washingtonpost.com/news/the-switch/wp/2018/05/22/amazon-is-selling-facial-recognition-to-law-enforcement-for-a-fistful-of-dollars/?noredirect=on&utm_term=.07d9ca13ab77.

[81] See, for example, J. Stanley, "FBI and Industry Failing to Provide Needed Protections for Face Recognition." *ACLU—Free Future*, June 15, 2016. Available: https://www.aclu.org/blog/privacy-technology/surveillance-technologies/fbi-and-industry-failing-provide-needed.

[82] It is also the case that, among the false positives, nonwhite members of Congress were overrepresented relative to their proportion in Congress as a whole, perhaps indicating that the accuracy of the technology is, to some degree, race-dependent. Without knowing more about the composition of the mugshot database, however, we cannot assess the significance of this result.

[83] See J. Snow, "Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots." *ACLU—Free Future*, July 26, 2018. Available: https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28. See also R. Brandom, "Amazon's facial recognition matched 28 members of Congress to criminal mugshots." *The Verge*, July 26, 2018. Available: https://www.theverge.com/2018/7/26/17615634/amazon-rekognition-aclu-mug-shot-congress-facial-recognition.

[84] See "Amazon Rekognition Developer Guide." Amazon, p. 131, 2018. Available: https://docs.aws.amazon.com/rekognition/latest/dg/rekognition-dg.pdf. Also see K. Tenbarge, "Amazon Responds to ACLU's Highly Critical Report of Rekognition Tech," Inverse, July 26, 2018. Available: https://www.yahoo.com/news/amazon-responds-aclu-apos-highly-160000264.html.

[85] The story also highlights the question of accountability, illustrating how the principles discussed in this report intersect with and complement each other.

Law

[86] Of course, competent use does not preclude use for bad ends (e.g., government surveillance that impinges on human rights). The principle of competence is one principle in a set that, collectively, is designed to ensure the ethical application of A/IS. See the EAD chapter "General Principles".

[87] Developing "well grounded" guidelines will typically require that the creators of A/IS gather input from both those operating the technology and those affected by the technology's operation.

[88] The use of facial recognition technologies by security and law enforcement agencies raises issues that extend beyond the question of operator competence. For further discussion of such issues, see C. Garvie, A. M. Bedoya, and J. Frankle, "The Perpetual Line-Up: Unregulated Police Face Recognition in America," *Georgetown Law, Center on Privacy & Technology*, October 18, 2016, Available: https://www.perpetuallineup.org/.

[89] As noted above, some professional organizations, such as the ABA, have begun to recognize in their codes of ethics the importance of technological competence, although the guidance does not yet address A/IS specifically.

[90] Including those engaged in the procurement and deployment of a system means that those acquiring and authorizing the use of a system can share in the responsibility for its results. For example, in the case of A/IS deployed in the service of the courts, this could be the judiciary; in the case of A/IS deployed in the service of law enforcement, this could be the agency responsible for the enforcement of the law and the administration of justice; in the case of A/IS used by a party to legal proceedings, this could be the party's counsel.

[91] J. New and D. Castro, "How Policymakers Can Foster Algorithmic Accountability."*Information Technology & Innovation Foundation,* p. 5, 2018. Available: https://www.itif.org/publications/2018/05/21/how-policymakers-can-foster-algorithmic-accountability.

[92] Included among possible "causes" for an effect are not only the decision-making pathways of algorithms but also, importantly, the decisions made by humans involved in the design, development, procurement, deployment, operation, and validation of effectiveness of A/IS.

[93] The challenge, moreover, is one not only of assigning responsibility, but of assigning levels of responsibility (a task that could benefit from a neutral model that could consider how much interaction and influence each stakeholder has in every decision).

[94] Scherer (2016): 372. In addition to diffuseness, Scherer identifies discreetness, discreteness, and opacity as features of the design and development of A/IS that make apportioning responsibility for their outcomes a challenge for regulators and courts.

[95] In answering these questions, it will be important to keep in mind the distinction between responsibility (a factual question) and ultimate accountability (a normative question). In the case of the example under discussion, there may be multiple individuals who have

Law

some practical responsibility for the sentence given, but the normative framework may place ultimate accountability on the judge. Before normative accountability can be assigned, however, pragmatic responsibilities must be clarified and understood. Hence the focus, in this section, on clarifying lines of responsibility so that ultimate accountability can be determined.

[96] If effectiveness is measured against statistics that themselves may represent human bias (e.g., arrest rates), then the effectiveness measures may just reflect and reinforce that bias.

[97] "'The algorithm did it' is not an acceptable excuse if algorithmic systems make mistakes or have undesired consequences, including from machine-learning processes." See "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms." FAT/ML Resources. www.fatml.org/resources/principles-for-accountable-algorithms.

[98] See Langewiesche, W. 1998. "The Lessons of ValuJet 592". *Atlantic Monthly*. 281: 81-97; S. D. Sagan. *Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton University Press, 1995.

[99] For a discussion of the role of explanation in maintaining accountability for the results of A/IS and of the question of whether the standards for explanation should be different for A/IS than they are for humans, see F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. J. Gershman, D. O'Brien, S. Shieber, J. Waldo, D. Weinberger, and A. Wood, Accountability of AI Under the Law: The Role of Explanation (November 3, 2017). Berkman Center Research Publication Forthcoming; Harvard Public Law Working Paper No. 18-07. Available: https://ssrn.com/abstract=3064761 or http://dx.doi.org/10.2139/ssrn.3064761.

[100] Also, gaining access to that information should not be unduly burdensome.

[101] Those developing a model for accountability for A/IS may find helpful guidance in considering models of accountability used in other domains (e.g., data protection).

[102] For a discussion of how such policies might be implemented in accordance with protocols for information governance, see J. R. Baron and K. E. Armstrong, "The Algorithm in the C-Suite: Applying Lessons Learned and Information Governance Best Practices to Achieve Greater Post-GDPR Algorithmic Accountability," in *The GDPR Challenge: Privacy, Technology, and Compliance In An Age of Accelerating Change*, A. Taal, Ed. Boca Raton, FL: CRC Press, forthcoming.

[103] These inquiries can be supported by technological tools that may provide information essential to answering questions of accountability but that do not require full transparency into underlying computer code and may avoid the necessity of an intrusive audit; see Kroll et al. (2017). Among the tools identified by Kroll and his colleagues are: software verification, cryptographic commitments, zero-knowledge proofs, and fair random choices. While the use of such tools may avoid the limitations of solutions such as transparency and audit, they do require that creators of A/IS design their systems so that they will be compatible with the application of such tests.

Law

[104] Certifications may include, for example, professional certifications of competence, but also certifications of compliance of processes with standards. An example of a certification program specifically addressing A/IS is *The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)*, https://standards.ieee.org/industry-connections/ecpais.html.

[105] This means that A/IS used in legal systems will have to be defensible in courts. The margin of error will have to be low or the use of A/IS will not be permitted.

[106] It is also the case that evidence produced by A/IS will be subject to chain-of-custody rules, as are other types of forensic evidence, to ensure integrity, confidentiality, and authenticity.

[107] See for instance Art. 22(1) Regulation (EU) 2016/679.

[108] Human dignity, as a core value protected by the United Nations Universal Declaration of Human Rights, requires us to fully respect the personality of each human being and prohibits their objectification.

[109] This concern is reflected in Principle 5 of the European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment, recently published by the Council of Europe's European Commission for the Efficiency of Justice (CEPEJ). Principle 5 ("Principle 'Under User Control': preclude a prescriptive approach and ensure that users are informed actors and in control of the choices made") states, with regard to professionals in the justice system that they should "at any moment, be able to review judicial decisions and the data used to produce a result and continue not to be necessarily bound by it in the light of the specific features of that particular case," and, with regard to decision subjects, that he or she must "be clearly informed of any prior processing of a case by artificial intelligence before or during a judicial process and have the right to object, so that his/her case can be heard directly by a court." See CEPEJ, *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment* (Strasbourg, 2018), p. 10.

[110] J. Tashea, Calculating Crime: Attorneys are Challenging the Use of Algorithms to Help Determine Bail, Sentencing and Parole, ABA Journal (March 2017).

[111] Loomis v. Wisconsin, 68 WI. (2016).

[112] *Id.* at pp. 46-66.

[113] R. Wexler, Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System, Stanford Law Review, 2018.

[114] Malenchik v. State, 928 N.E.2d 564, 574 (Ind. 2010).

[115] People *v.* Chubbs CA2/4, B258569 (Cal. Ct. App. 2015).

[116] U.S. v. Ocasio, No. 3:11-cr-02728-KC, slip op. at 1-2, 11-12 (W.D. Tex. May 28, 2013).

[117] U.S. v. Johnson, No. 1:15-cr-00565-VEC, order (S.D.N.Y., June 7, 2016).

[118] Indeed, without transparency, there may, in some circumstances, be no means for even knowing whether an error that needs to be corrected was committed. In the case of A/IS

Law

applied in a legal system, an "error" can mean real harm to the dignity, liberty, and life of an individual.

[119] *Fairness* (as well as *bias*) can be defined in more than one way. For purposes of this discussion, a commitment is not made to any one definition—and indeed, it may not be either desirable or feasible to arrive at a single definition that would be applied in all circumstances. For purposes of this discussion, the key point is that transparency will be essential in building informed trust in the fairness of a system, regardless of the specific definition of *fairness* that is operative.

[120] To the extent permitted by the normal operation of the A/IS: allowing for, for example, variation in the human inputs to a system that may not be eliminated in any attempt at replication.

[121] With regard to information explaining how a system arrived at a given output, GDPR makes provision for a decision subject's right to an explanation of algorithmic decisions affecting him or her: automated processing of personal data "should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision." GDPR, Recital 71.

[122] Even among sensitive data, some data may be more sensitive than others. See I. Ajunwa, "Genetic Testing Meets Big Data: Tort and Contract Law Issues," 75 *Ohio St. L. J.* 1225 (2014). Available: https://ssrn.com/abstract=2460891.

[123] See A. Baker, "Updated N.Y.P.D. Anti-Crime System to Ask: 'How We Doing?'" *New York Times*, May 8, 2017, https://www.nytimes.com/2017/05/08/nyregion/nypd-compstat-crime-mapping.html; S. Weichselbaum, "How a 'Sentiment Meter' Helps Cops Understand Their Precincts," *Wired*, July 16, 2018. Available: https://www.wired.com/story/elucd-sentiment-meter-helps-cops-understand-precincts/.

[124] This table is a preliminary draft and is meant only to illustrate a useful tool for facilitating reasoning about who should have access to what information. Other categories of stakeholder and other categories of information (e.g., the identity and nature of the designer/manufacturer of the A/IS, the identity and nature of the investors backing a particular system or company) could be added as needed.

[125] For discussions of these two dimensions of explanation, see S. Wachter, et al. (2017). "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation"; A. Selbst, and S. Barocas, The Intuitive Appeal of Explainable Machines.

[126] Wexler, Rebecca. 2018. "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System". *Stanford Law Review*. 70 (5): 1342-1429; Tashea, Jason. "Federal judge releases DNA software source code that was used by New York City's crime lab." *ABA Journal* (2017). http://www.abajournal.com/news/article/federal_judge_releases_dna_software_source_code.

[127] Or, if two approaches are found to be, for practical purposes, equally effective, the simpler, more easily explained approach may be preferred.

Law

[128] For a discussion of the limits of transparency and of alternative modes of gaining actionable answers to questions of verification and accountability, see J.A. Kroll, J. Huey, S. Barocas, E.W. Felten, J.R. Reidenberg, D.G. Robinson, H. Yu, "Accountable Algorithms" (March 2, 2016). *University of Pennsylvania Law Review*, Vol. 165, 2017 Forthcoming; Fordham Law Legal Studies Research Paper No. 2765268. Available at SSRN: https://ssrn.com/abstract=2765268. See also J.A. Kroll, The fallacy of inscrutability, *Phil. Trans. R. Soc.* A 376: 20180084. http://dx.doi.org/10.1098/rsta.2018.0084 (Note p. 9: "While transparency is often taken to mean the disclosure of source code or data, possibly to a trusted entity such as a regulator, this is neither necessary nor sufficient for improving understanding of a system, and it does not capture the full meaning of transparency.")

[129] In particular with respect to due process, the current dialogue on the use of A/IS centers on the tension between the need for transparency and the need for the protection of intellectual property rights. Adhering to the principle of Effectiveness as articulated in this work can substantially help in defusing this tension. Reliable empirical evidence of the effectiveness of A/IS in meeting specific real-world objectives may foster informed trust in such A/IS, without disclosure of proprietary or trade secret information.

[130] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," SSRN Electronic Journal, p. 5, 2017 for the example cited.

[131] W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood, "Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations," The RAND Corporation, pp. 67-69, 2013.

[132] Support from the University of Memphis was led by Richard Janikowski, founding Director of the Center for Community Criminology and Research (School of Urban Affairs and Public Policy, the University of Memphis) and the Shared Urban Data System (The University of Memphis).

[133] E. Figg, "The Legacy of Blue CRUSH," High Ground, March 19, 2014.

[134] Figg, "Legacy."

[135] Nucleus Research, *ROI Case Study: IBM SPSS—Memphis Police Department*, Boston, Mass., Document K31, June 2010. Perry et al., *Predictive Policing*, 69.

[136] Figg, "Legacy."

[137] Figg, "Legacy."

[138] See: AI Now, *Algorithmic Accountability Policy Toolkit*, p. 12, Oct. 2018. Available: https://ainowinstitute.org/aap-toolkit.pdf; D. Robinson and L. Koepke, *Stuck in a Pattern: Early evidence on "predictive policing" and civil rights*, Upturn, Aug. 2016. Available: https://www.upturn.org/reports/2016/stuck-in-a-pattern/; S. Brayne, "Big Data Surveillance: The Case of Policing," American Sociological Review, 2016. Available: https://journals.sagepub.com/doi/10.1177/0003122417725865; A. G. Ferguson, "Policing Predictive Policing,"

Law

Washington University Law Review, vol. 94, no. 5, 2017. Available: https://openscholarship.wustl.edu/law_lawreview/vol94/iss5/5/; K. Lum and W. Isaac, "To predict and serve?" *Significance* 2016. Available: https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1740-9713.2016.00960.x; B. J. Jefferson, "Predictable Policing: Predictive Crime Mapping and Geographies of Policing and Race," Annals of the American Association of Geographers, vol. 108, no. 1, pp. 1-16, 2018. Available: https://doi.org/10.1080/24694452.2017.1293500.

[139] For a discussion of the criteria that may define a "high-crime area," and so potentially license more intrusive policing, see A. G. Ferguson and D. Bernache, "The 'High-Crime Area' Question: Requiring Verifiable and Quantifiable Evidence for Fourth Amendment Reasonable Suspicion Analysis," American University Law Review, vol. 57, pp. 1587-1644.

[140] While A/IS, if misapplied, may perpetuate bias, it holds at least the potential, if applied with appropriate controls, to reduce bias. For a study of how an impersonal technology such as a red light camera may reduce bias, see R. J. Eger, C. K. Fortner, and C. P. Slade, "The Policy of Enforcement: Red Light Cameras and Racial Profiling," *Police Quarterly*, pp. 1-17, 2015. Available: http://hdl.handle.net/10945/46909.

[141] See, for example: J. Tashea, "Estonia considering new legal status for artificial intelligence," ABA Journal, Oct. 20, 2017, and European Parliament Resolution of Feb. 16, 2017.

[142] See Legal Entity, Person, *in* B. Bryan A. Garner, Black's Law Dictionary, 10th Edition. Thomas West, 2014.

[143] J. S. Nelson, "Paper Dragon Thieves." Georgetown Law Journal 105 (2017): 871-941.

[144] M. U. Scherer, "Of Wild Beasts and Digital Analogues: The Legal Status of Autonomous Systems." Nevada Law Journal 19, forthcoming 2018.

[145] See M. U. Scherer, "Of Wild Beasts and Digital Analogues: The Legal Status of Autonomous Systems." Nevada Law Journal 19, forthcoming 2018; J. F. Weaver. Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws. Santa Barbara, CA: Praeger, 2013; L. B. Solum. "Legal Personhood for Artificial Intelligences." North Carolina Law Review 70, no. 4 (1992): 1231–1287.

About *Ethically Aligned Design*

The Mission and Results 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.

To advance toward this goal, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems brought together more than a thousand 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 surrounding autonomous and intelligent systems.

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

About *Ethically Aligned Design*

From Principles to Practice– Results from our Work to Date

In addition to the creation of *Ethically Aligned Design*, The IEEE Global Initiative, independently or through the IEEE Standards Association, has directly inspired the following works:

- **The launch of the IEEE P7000™ series of approved standardization projects**

  This is the first series of standards in the history of the IEEE Standards Association that explicitly focuses on societal and ethical issues associated with a certain field of technology

  More information can be found at: ethicsinaction.ieee.org

- **Artificial Intelligence and Ethics in Design**

  These ten courses are designed for global professionals, as well as their managers, working in engineering, IT, computer science, big data, artificial intelligence, and related fields across all industries who require up-to-date information on the latest technologies. The courses explicitly mirror content from *Ethically Aligned Design*, and feature numerous experts as instructors who helped create *Ethically Aligned Design*.

  More information can be found at: innovationatwork.ieee.org/courses/artificial-intelligence-and-ethics-in-design

- **The creation of an A/IS Ethics Glossary**

  The Glossary features more than two hundred pages of terms that help to define the context of A/IS ethics for multiple stakeholder groups, specifically: engineers, policy makers, philosophers, standards developers, and computational disciplines experts. It is currently in its second iteration and has also been informed by the IEEE P7000™ standards working groups.

  Download the Glossary at: standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e_glossary.pdf

- **The launch of OCEANIS**

  The IEEE Standards Association, inspired by the work of The IEEE Global Initiative, has contributed significantly to the establishment of The Open Community for Ethics in Autonomous and Intelligent Systems (OCEANIS). It is a global forum for discussion, debate, and collaboration for organizations interested in the development and use of standards to further the creation of autonomous and intelligent systems. OCEANIS members are working together to enhance the understanding of the role of standards in facilitating innovation, while addressing problems that expand beyond technical solutions to addressing ethics and values.

  More information can be found at: ethicsstandards.org

About *Ethically Aligned Design*

- **The launch of ECPAIS**

  The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) has the goal to create specifications for certification and marking processes that advance transparency, accountability, and reduction in algorithmic bias in autonomous and intelligent systems. ECPAIS intends to offer a process and define a series of marks by which organizations can seek certifications for their processes around the A/IS products, systems, and services they provide.

  More information can be found at: standards.ieee.org/industry-connections/ecpais.html

- **The launch of CXI**

  The Council on Extended Intelligence (CXI) was directly inspired by the work of The IEEE Global Initiative and the work of The MIT Media Lab around "Extended Intelligence". CXI was launched jointly by the IEEE Standards Association and The MIT Media Lab. CXI's mission is to proliferate the ideals of responsible participant design, data agency, and metrics of economic prosperity, prioritizing people and the planet over profit and productivity. Membership includes thought leaders from the EU Parliament and Commission, the UK House of Lords, the OECD, the United Nations, local and national administrations, and renowned experts in economics, data science, and multiple other disciplines from around the world.

  More information can be found at: globalcxi.org

- **The launch of EADUC**

  The Ethically Aligned Design University Consortium (EADUC) is being established with the aim to reach every engineer at the beginning of their studies to help them prioritize values-driven, applied ethical principles at the core of their work. Working in conjunction with philosophers, designers, social scientists, academics, data scientists, and the corporate and policy communities, EADUC also has the goal that *Ethically Aligned Design* will be used in teaching at all levels of education globally as the new vision for design in the algorithmic age.

- **The launch of "AI Commons"**

  The work of The IEEE Global Initiative has delivered key ideas and inspiration that are rapidly evolving toward establishing a global collaborative platform around A/IS. The mission of AI Commons is to gather a true ecosystem to democratize access to AI capabilities and thus to allow anyone, anywhere to benefit from the possibilities that AI can provide. In addition, the group will be working to connect problem owners with the community of solvers, to collectively create solutions with AI. The ultimate goal is to implement a framework for participation and cooperation to make using and benefiting from AI available to all.

  More information can be found at: www.aicommons.com

About *Ethically Aligned Design*

IEEE P7000™ Approved Standardization Projects

The IEEE P7000™ series of standards projects under development represents a unique addition to the collection of over 1,900 global IEEE standards and projects. Whereas more traditional standards have a focus on technology interoperability, functionality, safety, and trade facilitation, the IEEE P7000 series addresses specific issues at the intersection of technological and ethical considerations. Like its technical standards counterparts, the IEEE P7000 series empowers innovation across borders and enables societal benefit.

For more information or to join any working group, please see the links below. Committees that authored *Ethically Aligned Design,* as well as other committees within IEEE, that created specific working groups are listed below each project.

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

  Inspired by Methodologies to Guide Ethical Research and Design Committee, and supported by IEEE Computer Society
  standards.ieee.org/project/7000.html

- **IEEE P7001™ -** IEEE Standards Project for Transparency of Autonomous Systems

  Inspired by the General Principles Committee, and supported by IEEE Vehicular Technology Society
  standards.ieee.org/project/7001.html

- **IEEE P7002™ -** IEEE Standards Project for Data Privacy Process

  Inspired by The Personal Data and Individual Agency Control Committee, and supported by IEEE Computer Society
  standards.ieee.org/project/7002.html

- **IEEE P7003™ -** IEEE Standards Project for Algorithmic Bias Considerations

  Supported by IEEE Computer Society
  standards.ieee.org/project/7003.html

- **IEEE P7004™ -** IEEE Standards Project for Child and Student Data Governance

  Inspired by The Personal Data and Individual Agency Control Committee, and supported by IEEE Computer Society
  standards.ieee.org/project/7004.html

About *Ethically Aligned Design*

- **IEEE P7005™ -** IEEE Standards Project for Employer Data Governance

  Inspired by The Personal Data and Individual Agency Control Committee, and supported by IEEE Computer Society
  standards.ieee.org/project/7005.html

- **IEEE P7006™ -** IEEE Standards Project for Personal Data AI Agent Working Group

  Inspired by The Personal Data and Individual Agency Control Committee, and supported by IEEE Computer Society
  standards.ieee.org/project/7006.html

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

  Supported by IEEE Robotics and Automation Society
  standards.ieee.org/project/7007.html

- **IEEE P7008™ -** IEEE Standards Project for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems

  Inspired by the Affective Computing Committee, and supported by IEEE Robotics and Automation Society
  standards.ieee.org/project/7008.html

- **IEEE P7009™** - IEEE Standards Project for Fail-Safe Design of Autonomous and Semi-Autonomous Systems

  Supported by IEEE Reliability Society
  standards.ieee.org/project/7009.html

- **IEEE P7010™ -** IEEE Standards Project for Well-being Metric for Autonomous and Intelligent Systems

  Inspired by the Well-being Committee, and supported by IEEE Systems, Man and Cybernetics Society
  standards.ieee.org/project/7010.html

- **IEEE P7011™ -** IEEE Standards Project for the Process of Identifying and Rating the Trustworthiness of News Sources

  Supported by IEEE Society on Social Implications of Technology
  standards.ieee.org/project/7011.html

- **IEEE P7012™ -** IEEE Standards Project for Machine Readable Personal Privacy Terms

  Supported by IEEE Society on Social Implications of Technology
  standards.ieee.org/project/7012.html

- **IEEE P7013™ -** IEEE Standards Project for Inclusion and Application Standards for Automated Facial Analysis Technology

  Supported by IEEE Society on Social Implications of Technology
  standards.ieee.org/project/7013.html

About *Ethically Aligned Design*

Who We Are

About IEEE

IEEE is the largest technical professional organization dedicated to advancing technology for the benefit of humanity, with over 420,000 members in more than 160 countries. Through its highly cited publications, conferences, technology standards, and professional and educational activities, IEEE is the trusted voice in a wide variety of areas ranging from aerospace systems, computers, and telecommunications to biomedical engineering, electric power, and consumer electronics.

To learn more, visit the IEEE website: www.ieee.org

About the IEEE Standards Association

The IEEE Standards Association (IEEE-SA), a globally recognized standards-setting body within IEEE, develops consensus standards through an open process that engages industry and brings together a broad stakeholder community. IEEE standards set specifications and best practices based on current scientific and technological knowledge. The IEEE-SA has a portfolio of over 1,900 active standards and over 650 standards under development.

For more information, visit the IEEE-SA website: standards.ieee.org

About The IEEE Global Initiative

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

To learn more, visit The IEEE Global Initiative website:
standards.ieee.org/industry-connections/ec/autonomous-systems.html

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.

Names of experts involved in the various committees of The IEEE Global Initiative can be found at: standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf

IEEE makes all versions of *Ethically Aligned Design* available under the Creative Commons Attribution-Non-Commercial 4.0 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 *Ethically Aligned Design* content and subject matter will be selected for submission into formal IEEE processes, including standards development and education purposes.

The IEEE Global Initiative and *Ethically Aligned Design* contribute, together with other efforts within IEEE, such as IEEE TechEthics™, (techethics.ieee.org), to a broader effort at IEEE to foster open, broad, and inclusive conversation about ethics in technology.

About *Ethically Aligned Design*

Our Process

To ensure the greatest cultural relevance and intellectual rigor possible in our work, The IEEE Global Initiative has sought for and received global feedback for versions 1 and 2 (after hundreds of experts created first drafts) to inform this *Ethically Aligned Design*, *First Edition (EAD1e).*

We released ***Ethically Aligned Design,* Version 1** *(EADv1)* as a Request for Input in December of 2016 and received **over two hundred pages** of in-depth feedback about the draft. We subsequently released ***Ethically Aligned Design,* Version 2** *(EADv2)* in December 2017 and received **over three hundred pages** of in-depth feedback about the draft. This feedback included further insights about the eight original sections from EADv1, along with unique/new input for the five new sections included in EADv2.

Both versions included "candidate recommendations" instead of direct "recommendations", because our communities had been engaged in debate and weighing various options.

This process was taken to the next level with *Ethically Aligned Design, First Edition (EAD1e),* using EADv1 and EADv2 as its initial foundation. Although we expect future editions of *Ethically Aligned Design*, a vetting process has taken place within the global community that gave rise to this seminal work. Therefore, we can now speak of "recommendations" without any further restriction, and *EAD1e* also includes a set of policy recommendations.

This process included matters of "internal consistency" across the various chapters of *EAD1e* and also more specific or broader criteria, such as maturity of the specific chapters and consistency with respect to policy statements of IEEE. The review also considered the need for IEEE to maintain a neutral ~~a~~nd thus credible—position in areas and processes where it is likely that IEEE may become active in the future.

Beyond these formal procedures, the Board of Governors of IEEE Standards Association has endorsed the work of the IEEE Global Initiative and offers it for consideration by governments, businesses, and the public at large with the following resolution:

Whereas the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems is an authorized activity within the IEEE Standards Association Industry Connections program created with the stated mission:

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;

Whereas versions 1 and 2 of *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS)* were developed as calls for comment and candidate recommendations by several hundred professionals including engineers, scientists,

About *Ethically Aligned Design*

ethicists, sociologists, economists, and many others from six continents;

Whereas the recommendations contained in *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), First Edition* are the result of the consideration of hundreds of comments submitted by professionals and the public at large on versions 1 and 2;

Whereas through an extensive, global, and open collaborative process, more than a thousand experts of The IEEE Global Initiative have developed and are in the process of final editing and publishing *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), First Edition*; now, therefore, be it

Resolved, that the IEEE Standards Association Board of Governors:

1. expresses its appreciation to the leadership and members of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems for the creation of *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), First Edition;* and
2. supports and commends the collaborative process used by The IEEE Global Initiative to achieve extraordinary consensus in such complex and vast matters in less than three years; and
3. endorses and offers *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), First Edition* to businesses, governments and the public at large for consideration and guidance in the ethical development of autonomous and intelligent systems.

Terminology Update

For *Ethically Aligned Design*, we prefer not to use—as far as possible—the vague term "AI" and use instead the term, *autonomous and intelligent systems* (or *A/IS*). Even so, it is inherently difficult to define "intelligence" and "autonomy". One could, however, limit the scope for practical purposes to computational systems using algorithms and data to address complex problems and situations, including the capability of improving their performance based on evaluating previous decisions, and say that such systems could be considered as "intelligent".

Such systems could be regarded also as "autonomous" in a given domain as long as they are capable of accomplishing their tasks despite environment changes within the given domain. This terminology is applied throughout *Ethically Aligned Design*, *First Edition* to ensure the broadest possible application of ethical considerations in the design of the addressed technologies and systems.

About *Ethically Aligned Design*

How the Document Was Prepared

This document was developed by The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, which is an authorized Industry Connections activity within the IEEE Standards Association, a Major Organizational Unit of IEEE.

It was prepared using an open, collaborative, and consensus building approach, following the processes of the **Industry Connections framework program** of the IEEE Standards Association **(standards.ieee.org/industry-connections)**. This process does not necessarily incorporate all comments or reflect the views of every contributor listed in the Acknowledgements above or after each chapter of this work.

The views and opinions expressed in this collaborative work are those of the authors and do not necessarily reflect the official policy or position of their respective institutions or of the Institute of Electrical and Electronics Engineers (IEEE). This work is published under the auspices of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems for the purposes of furthering public understanding of the importance of addressing ethical considerations in the design of autonomous and intelligent systems.

In no event shall IEEE or IEEE-SA Industry Connections Activity Members be liable for any errors, omissions or damage, direct or otherwise, however caused, arising in any way out of the use of or application of any recommendation contained in this publication.

The Board of Governors of the IEEE Standards Association, its highest governing body, commends the consensus-building process used in developing *Ethically Aligned Design*, First Edition, and offers the work for consideration and guidance in the ethical development of autonomous and intelligent systems.

How to Cite *Ethically Aligned Design*

Please cite *Ethically Aligned Design, First Edition* 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*, First Edition. IEEE, 2019. **https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html**

Key References

Key reference documents listed in *Ethically Aligned Design*, First Edition:

- **Appendix 1 -** The State of Well-being Metrics (An Introduction)
  bit.ly/ead1e-appendix1
  (Referenced in Well-being Section)
- **Appendix 2 -** The Happiness Screening Tool for Business Product Decisions
  bit.ly/ead1e-appendix2
  (Referenced in Well-being Section)
- **Appendix 3 -** Additional Resources: Standards Development Models and Frameworks
  bit.ly/ead1e-appendix3
  (Referenced in Well-being Section)
- **Glossary**
  bit.ly/ead1e-glossary

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, Incorporated ("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 and Ethically Aligned Design contribute to broader efforts at IEEE about ethics in technology.

IEEE
Advancing Technology for Humanity

082-03-19