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*.