

General Principles

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

1. Embody the highest ideals of human beneficence as a superset of Human Rights.
2. Prioritize benefits to humanity and the natural environment from the use of A/IS. Note that these should not be at odds – one depends on the other. Prioritizing human well-being does not mean degrading the environment.
3. Mitigate risks and negative impacts, including misuse, as A/IS evolve as socio-technical systems. In particular by ensuring A/IS are accountable and transparent.

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

We have identified principles created by our Committee as well as aggregated principles reflected from other Committees of The IEEE Global Initiative. Therefore, readers should note that some general principles are reiterated and elaborated by other committees, as appropriate to the specific concerns of those committees. We have purposefully structured our Committee and this document in this way to provide readers with a broad sense of the themes and ideals reflecting the nature of ethical alignment for these technologies as an introduction to our overall mission and work.

General Principles

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

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

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

General Principles

Principle 1 – Human Rights

Issue:

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

Background

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

1. A/IS should be designed and operated in a way that both respects and fulfills human rights, freedoms, human dignity, and cultural diversity.
2. A/IS must be verifiably safe and secure throughout their operational lifetime.

3. If an A/IS causes harm it must always be possible to discover the root cause, by assuring *traceability* for said harm (see also Principle 4 – Transparency).

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

Candidate Recommendations

To best honor human rights, society must assure the safety and security of A/IS so that they are designed and operated in a way that benefits humans:

1. Governance frameworks, including standards and regulatory bodies, should be established to oversee processes assuring that the

General Principles

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

2. A way to translate existing and forthcoming legal obligations into informed policy and technical considerations is needed. Such a method should allow for differing cultural norms as well as legal and regulatory frameworks.
3. For the foreseeable future, A/IS should not be granted rights and privileges equal to human rights: A/IS should always be subordinate to human judgment and control.

Further Resources

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

- [The Universal Declaration of Human Rights](#), 1947.
- [The International Covenant on Civil and Political Rights](#), 1966.
- [The International Covenant on Economic, Social and Cultural Rights](#), 1966.
- [The International Convention on the Elimination of All Forms of Racial Discrimination](#), 1965.
- [The Convention on the Rights of the Child](#).
- [The Convention on the Elimination of All Forms of Discrimination against Women](#), 1979.
- [The Convention on the Rights of Persons with Disabilities](#), 2006.
- [The Geneva Conventions and additional protocols](#), 1949.
- [IRTF's Research into Human Rights Protocol Considerations](#).
- [The UN Guiding Principles on Business and Human Rights](#), 2011.
- For an example of a guide on how to conduct an ethical risk assessment see British Standards Institute BS8611:2016, [Guide to the Ethical Design and Application of Robots and Robotic Systems](#).

General Principles

Principle 2 – Prioritizing Well-being

Issue:

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

Background

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

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

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

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

General Principles

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

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

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

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

fiscal metrics of success, we also risk expediting negative and irreversible harms to our planet and population.

Candidate Recommendation

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

Further Resources

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

General Principles

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

General Principles

Principle 3 – Accountability

Issue:

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

Background

The programming, output, and purpose of A/IS are often not discernible by the general public. Based on the cultural context, application, and use of A/IS, people and institutions need clarity around the manufacture and deployment of these systems to establish responsibility and accountability, and avoid potential harm. Additionally, manufacturers of these systems must be able to provide programmatic-level accountability proving why a system operates in certain ways to address legal issues of culpability, if necessary apportion culpability among several responsible designers, manufacturers, owners, and/or operators, to avoid confusion or fear within the general public.

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

Candidate Recommendations

To best address issues of responsibility and accountability:

1. Legislatures/courts should clarify issues of responsibility, culpability, liability, and accountability for A/IS where possible during development and deployment (so that manufacturers and users understand their rights and obligations).
2. Designers and developers of A/IS should remain aware of, and take into account when relevant, the diversity of existing cultural norms among the groups of users of these A/IS.
3. Multi-stakeholder ecosystems should be developed to help create norms (which can mature to best practices and laws) where they do not exist because A/IS-oriented technology and their impacts are too new (including representatives of civil society, law enforcement, insurers, manufacturers, engineers, lawyers, etc.).
4. Systems for registration and record-keeping should be created so that it is always possible to find out who is legally responsible for a particular A/IS. Manufacturers/operators/

General Principles

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

- Intended use
- Training data/training environment (if applicable)
- Sensors/real world data sources
- Algorithms
- Process graphs
- Model features (at various levels)
- User interfaces
- Actuators/outputs
- Optimization goal/loss function/reward function

Further Resources

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

General Principles

Principle 4 – Transparency

Issue: How can we ensure that A/IS are transparent?

Background

A key concern over autonomous systems is that their operation must be transparent to a wide range of stakeholders for different reasons (noting that the level of transparency will necessarily be different for each stakeholder). Stated simply, transparent A/IS are ones in which it is possible to discover how and why a system made a particular decision, or in the case of a robot, acted the way it did. Note that here the term transparency also addresses the concepts of traceability, explicability, and interpretability.

A/IS will be performing tasks that are far more complex and have more effect on our world than prior generations of technology. This reality will be particularly acute with systems that interact with the physical world, thus raising the potential level of harm that such a system could cause. For example, some A/IS already have real consequences to human safety or well-being, such as medical diagnosis AI systems, or driverless car autopilots; systems such as these are *safety-critical* systems.

At the same time, the complexity of A/IS technology will make it difficult for users of those systems to understand the capabilities and limitations of the AI systems that they use, or with which they interact. This opacity, combined with the often-decentralized manner in which it is developed, will complicate efforts to determine and allocate responsibility when something goes wrong with an AI system. Thus, lack of transparency both increases the risk and magnitude of harm (users not understanding the systems they are using) and also increases the difficulty of ensuring accountability (see Principle 3— Accountability).

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

1. For users, transparency is important because it provides a simple way for them to understand what the system is doing and why.
2. For validation and certification of an A/IS, transparency is important because it exposes the system's processes and input data to scrutiny.
3. If accidents occur, the AS will need to be transparent to an accident investigator, so the internal process that led to the accident can be understood.

General Principles

4. Following an accident, judges, juries, lawyers, and expert witnesses involved in the trial process require transparency to inform evidence and decision-making.
5. For disruptive technologies, such as driverless cars, a certain level of transparency to wider society is needed to build public confidence in the technology, promote safer practices, and facilitate wider societal adoption.

Candidate Recommendation

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

*Note that IEEE Standards [Working Group P7001™](#) has been set up in response to this recommendation.

Further Resources

- Cappelli, C., P. Engiel, R. Mendes de Araujo, and J. C. Sampaio do Prado Leite. "Managing Transparency Guided by a Maturity Model." *3rd Global Conference on Transparency Research* 1 no. 3, 1–17. Jouy-en-Josas, France: HEC Paris, 2013.
- Sampaio do Prado Leite, J. C., and C. Cappelli. "Software Transparency." *Business & Information Systems Engineering* 2, no. 3 (2010): 127–139.
- Winfield, A., and M. Jirotko. "The Case for an Ethical Black Box." *Lecture Notes in Artificial Intelligence* 10454, (2017): 262–273.
- Wortham, R. R., A. Theodorou, and J. J. Bryson. "What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems." *IJCAI-2016 Ethics for Artificial Intelligence Workshop*. New York, 2016.
- Machine Intelligence Research Institute. "[Transparency in Safety-Critical Systems](#)." August 25, 2013.
- Scherer, M. "[Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies](#)." *Harvard Journal of Law & Technology* 29, no. 2 (2015).
- U.K. House of Commons. "Decision Making Transparency" pp. 17–18 in [Report of the U.K. House of Commons Science and Technology Committee on Robotics and Artificial Intelligence](#), September 13, 2016.

General Principles

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

Issue:

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

Background

New technologies give rise to greater risk of misuse, and this is especially true for A/IS. A/IS increases the impact of risks such as hacking, the misuse of personal data, “gaming,” or exploitation (e.g., of vulnerable users by unscrupulous parties). These are not theoretical risks. Cases of A/IS hacking have already been widely reported, of [driverless cars](#) for example. The EU’s General Data Protection Regulation (GDPR) provides [measures to remedy the misuse](#) of personal data. The Microsoft [Tay AI chatbot](#) was famously gamed when it mimicked deliberately offensive users. In an age where these powerful tools are easily available, there is a need for new kind of education for citizens to be sensitized to risks associated with the misuse of A/IS.

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

They also have a role to play in guiding new technology proactively toward beneficial ends.

Candidate Recommendations

Raise public awareness around the issues of potential A/IS technology misuse in an informed and measured way by:

1. Providing ethics education and security awareness that sensitizes society to the potential risks of misuse of A/IS (e.g., by providing “data privacy” warnings that some smart devices will collect their user’s personal data).
2. Delivering this education in scalable and effective ways, beginning with those having the greatest credibility and impact that also minimize generalized (e.g., non-productive) fear about A/IS (e.g., via credible research institutions or think tanks via social media such as Facebook or YouTube).
3. Educating government, lawmakers, and enforcement agencies surrounding these issues so citizens work collaboratively with them to avoid fear or confusion (e.g., in the same way police officers have given public safety lectures in schools for years; in the near future they could provide workshops on safe A/IS).

General Principles

Further Resources

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