1 General Principles

The General Principles Committee seeks to articulate high-level ethical concerns that apply to all types of artificial intelligence and autonomous systems (AI/AS) regardless of whether they are physical robots (such as care robots or driverless cars) or software AIs (such as medical diagnosis systems, intelligent personal assistants, or algorithmic chat bots).

We are motivated by a desire to create ethical principles for AI/AS that:

1. Embody the highest ideals of human rights.

2. Prioritize the maximum benefit to humanity and the natural environment.

3. Mitigate risks and negative impacts as AI/AS evolve as socio-technical systems.

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

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

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

1 General Principles

Principle 1 – Human Benefit

Issue:

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

Background

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

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

Candidate Recommendations

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

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

Further Resources

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

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

1 General Principles

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

Principle 2 – Responsibility

Issue:

How can we assure that AI/AS are accountable?

Background

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

Candidate Recommendations

To best address issues of responsibility:

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

Further Resources

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

Principle 3 – Transparency

Issue:

How can we ensure that AI/AS are transparent?

Background

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

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

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

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

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

1 General Principles

Candidate Recommendation

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

Further Resources

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

Principle 4 – Education and Awareness

Issue:

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

Background

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

Candidate Recommendations

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

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

Further Resources

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