

Principles for Responsible AI

Review Version of 17 October 2018

Roger Clarke **

© Xamax Consultancy Pty Ltd, 2018

This document is at [{.html, .pdf}](http://www.rogerclarke.com/EC/PRAI)

1. Introduction

The last few years have seen another surge in interest and progress in AI. Several current technologies are widely considered to have the potential for significant impact. There is, however, nervousness among practitioners regarding the extent to which public concerns about negative impacts may hinder investment and deployment.

Journals, conferences and industry publications feature many discussions of the issues, a variety of ethical analyses, and proposals for 'principles for AI'. There are commonalities among the concerns and the proposals. However, even a cursory glance shows a great deal of diversity. Many elements are evident, yet most of them appear in only a small proportion of sources.

A comprehensive view of public expectations is highly desirable. In order to develop one, I identified a suite of 24 documents. Based on prior reading, I postulated a set of 10 themes. I then extracted propositions from the suite of documents, and allocated them to the themes. In order to achieve adequate cohesion and readability, I made adjustments to the themes, and to the expression of many of the elements.

The themes are presented in Table 2, identified as 'The 10 Principles', and the complete set of propositions is in Appendix 1, labelled 'The 50 Principles'. The remainder of this article briefly reviews AI, provides background to the process of extracting, organising and expressing the Principles, discusses their nature, and identifies potential applications.

2. What's 'AI'?

Some of the differences among the sources used in the present study might be expected to arise from varying interpretations of the nature of AI. Conventionally in the AI field, 'intelligence' is exhibited by an artefact if it evidences perception and cognition of (relevant aspects of) its environment, has goals, and formulates actions towards the achievement of those goals (Albus 1991, Russell & Norvig 2003, McCarthy 2007). The term 'artificial' implies 'human-made', using a yardstick that is open to interpretation variously as 'equivalent to human' or 'superior to human'.

In practice, a great deal of AI is 'different from human'. Given that artefacts and humans have comparative advantages over one another, it can be argued that a more

appropriate term would be 'complementary intelligence': "information technologists [should] delineate the relationship between robots and people by applying the concept of decision structuredness to blend computer-based and human elements advantageously" (Clarke 1989, 1993, 2014). 'Complementary intelligence' would (1) do things well that humans do badly or cannot do at all; and (2) function as elements within systems that include both humans and artefacts, with effective, efficient and adaptable interfacing among them all.

It is useful to distinguish three broad areas in which activities are currently attracting public attention. Robotics continues 'on the factory floor' and in warehouses. It provides low-level control over the attitude, position and course of craft on or in water and in the air. It is achieving market penetration in self-driving vehicles, variously on rails and otherwise, in controlled environments such as mines, quarries and dedicated bus routes, but recently also in more open environments. Further examples can be found in the 'Internet of Things' (IoT) movement, and related initiatives under such rubrics as 'smart houses' and 'smart cities'.

A second category of AI that I suggest needs to be considered is cyborgisation, by which I mean the process of enhancing individual humans by technological means, resulting in hybrids of a human and one or more artefacts (Clarke 2005, Warwick 2014). Many forms of cyborg fall outside the field of AI, such as spectacles, implanted lenses, stents, inert hip-replacements and SCUBA gear. However, a proportion of the artefacts used to enhance humans qualify, by combining sensors, computational or programmatic 'intelligence', and one or more actuators. Examples include heart pacemakers (since 1958), cochlear implants (since the 1960s, and commercially since 1978), and some replacement legs for above-knee amputees, in that the artificial knee contains software to sustain balance within the joint.

In the third area, data analytics, a key technique deriving from the machine learning (ML) area is neural networks. This differs from previous approaches, in that it does not necessarily begin with active and careful modelling of a real-world problem-solution, problem or even problem-domain. Rather than comprising a set of entities and relationships that mirrors the key elements and processes of a real-world system, a neural network model may be simply a list of input variables and a list of output variables (and, in the case of 'deep' networks, intermediary variables). The weightings imputed for each connection reflect the characteristics of the training-set that was fed in, and of the particular learning algorithm that was imposed on the training-set. These features combine with questions about the selectivity, accuracy and compatibility of the data to give rise to uncertainty about the technique's degree of affinity with the real world to which it is applied.

A common feature of these three areas is that, to a greater extent than in the past, software is drawing inferences, making decisions, and taking action. The field might be better characterised by adopting a term that reflects the change in emphasis. The term 'robotics' has been associated with the idea of 'machines that think'. An alternative term, such as 'intellectics', would have the advantage of instead implying 'computers that do'. Sensor-computer-actuator packages are now generating a strong impulse for action to be taken in and on the real world. The new world of intellectics sees artefacts at the very least communicating a recommendation to a human, but sometimes generating a default-decision that is subject to being countermanded or

overridden by a human, and even acting autonomously based on the inferences they have drawn.

In the near future, it may be worthwhile re-casting the propositions discussed below to address 'complementary intelligence' and 'intellectics'. Currently, however, the mainstream discussion is about 'AI', and the remainder of this document reflects that norm.

3. The AI Supply Chain

Robots act directly on the world, with varying degrees of autonomy. Cyborgisation involves interference with the human body, possibly consensually but possibly not. Not only do neural nets enable decision-making about humans by artefacts, but their decision-making is largely inscrutable. Many people reasonably feel discomfort about such developments, and some did even at the dawn of the robotics era: "every degree of independence we give the machine is a degree of possible defiance of our wishes. The genie in the bottle will not willingly go back in the bottle, nor have we any reason to expect them to be well disposed to us" (Wiener, 1949, quoted in Markoff 2013). Even technophiles might well anticipate that, if and when *homo sapiens* cedes power to *homo roboticus* and/or *roboticus sapiens*, it would involve a conscious act by humanity, rather than merely a long series of seemingly small decisions by technocrats (Menzel & D'Alusio 2001, Clarke 2014).

If the discomfort felt by people is to be addressed, and if such substantive problems as actually exist are to be solved, who bears what responsibilities? Discussion about responsibility for AI is often clouded by inadequate discrimination among the successive phases of the supply-chain from laboratory experiment to deployment in the field, and failure to assign responsibilities to the categories of entities that are active in each phase. Table 1 distinguishes among technology, artefacts that embody the technology, systems that incorporate the artefacts, the process of dissemination of artefacts and systems, and their application. In each phase, legal and moral responsibilities can then be assigned to researchers, to inventors, to innovators, to purveyors, and to users.

Table 1: Entities with Responsibilities in Relation to AI

<u>Phase</u>	<u>Result</u>	<u>Direct Responsibility</u>
Research	AI Technology	Researchers
Invention	AI-Based Artefacts	IR&D Engineers
Innovation	AI-Based Systems	Developers
Dissemination	Installed AI-Based Systems	Purveyors
Application	Impacts	User Organisations and Individuals

4. The Process

The 'Principles for Responsible AI' presented below arise from a consolidation of ideas from a suite of previously-published documents. The suite was assembled by surveying academic, professional and policy literatures. Diversity of perspective was actively sought. The sources include corporations and industry associations (5), governmental organisations (6), academics (4), professional associations (2), joint associations (2), and non-government organisations (5). Only sets that are available in the English language were used. An analysis by region of origin shows 4 pan-world, 3 pan-European, 10 US, 4 UK, 1 Australian, 1 Japanese and 1 Korean. Of the documents, 8 are formulations of 'ethical principles and IT', and the other 16 provide guidance specifically in relation to AI.

In the previous section and Table 1, distinctions were drawn among the responsible entities in the successive phases of the supply-chain. In only a few of the 24 documents in the suite were such distinctions evident, and in most cases it has to be interpolated which part of the supply-chain the document is intended to address. The European Parliament (CLA-EP 2016) refers to "design, implementation, dissemination and use", IEEE (2017) to "Manufacturers / operators / owners", GEFA (2016) to "manufacturers, programmers or operators", FLI (2017) to researchers, designers, developers and builders, and ACM (2017) to "Owners, designers, builders, users, and other stakeholders". Remarkably, however, in all of these cases the distinctions were only made within a single Principle rather than being applied to the set as a whole.

Some commonalities exist across the source documents. However, many of them contain only a few propositions, and overall there is far less consensus than might be expected 60 years after AI was first heralded. For example, only 1 document expressly encompasses cyborgisation (GEFA 2016); only 2 documents refer to the precautionary principle (CLA-EP 2016, GEFA 2016), and only 5 stipulate the conduct of impact assessments. One striking statistic is that only 3 of the c. 50 Principles are detectable in at least half of the documents in the set:

- ensure physical safety (17 / 24)
- ensure human control (12 / 24)
- ensure transparency of inferencing, decision-making and actions (12 / 24)

Each source naturally reflects the express, implicit and subliminal purposes of the drafters and the organisations on whose behalf they were composed. In some cases, for example, the set primarily addresses a specific form of AI, such as robotics or machine-learning. Documents prepared by corporations, industry associations, and even professional associations and joint associations tend to adopt the perspective of producer roles. For example, FLI (2017) perceives the need for "constructive and healthy exchange between AI researchers and policy-makers", but does not mention participation by stakeholders in the technology and its applications (at 3). As a result, transparency is constrained to a small sub-set of circumstances (at 6), 'designers and builders', rather than being responsible entities, are merely 'stakeholders in moral implications' (at 9), alignment with human values is seen as being necessary only in respect of "highly autonomous AI systems" (at 10), and "strict safety and control measures" are limited to a small sub-set of AI systems (at 22).

Similarly, ITIC (2017) considers that many responsibilities lie elsewhere, and assigns responsibilities to its members only in respect of safety, controllability and data quality. ACM (2017) is expressed in weak language (should be aware of, should encourage, are encouraged) and decision opaqueness is regarded as being acceptable, while IEEE (2017) suggests a range of important tasks for other parties (standards-setters, regulators, legislatures, courts), and phrases other suggestions in the passive voice, with the result that few obligations are clearly identified as falling on engineering professionals and the organisations that employ them. The House of Lords report might have been expected to adopt a societal or multi-stakeholder approach, yet, as reinforced by Smith (2018), it adopts the perspective of the AI industry.

The process of developing the Principles commenced with a set of themes derived from my prior background supplemented by first-pass reading of the selected documents. The documents were then inspected in greater detail. Propositions within each set were identified, extracted, and allocated to themes, maintaining back-references to the sources. Where items threw doubt on the structure or formulation of the general themes, the schema was adapted. Where similar points were expressed in varying ways, forms of words were selected or developed in order to sustain coherence and limit the extent to which the final set contains duplications. Of course, no claim is made that the selection of source-documents is complete or representative, or that the interpretations and the expressions used are the sole possibility, or even necessarily the most appropriate alternative.

5. The Principles

All themes, and all detailed propositions, have been expressed in imperative mode, i.e. in the form of instructions, in order to convey that they require action, rather than being merely desirable characteristics, or factors to be considered, or issues to be debated. In this form, the themes and propositions serve as a foundation for guiding and evaluating behaviour. They have therefore been referred to as 'Principles', in the sense of "a fundamental motive or reason for action, esp. one consciously recognized and followed" (OED 4a). The full set, in Appendix 1, is accordingly referred to as 'The 50 Principles', and the summary set, presented in Table 2, as 'The 10 Principles'.

In order to facilitate audit and re-analysis, access is provided to supporting materials (Clarke 2018b), comprising citations of, and extracts from, the 24 sources (Parts 1 and 2), a version of 'The 50 Principles' that includes back-references to the sources (Part 3), and a list of items that appear in source documents that have not been included in 'The 50 Principles', because they involve imprecise abstractions that are difficult to operationalise (e.g. 'human dignity', 'fairness' and 'justice'), or fall outside the scope of the present work (Part 4).

Each of the Principles requires somewhat different application in each phase of the AI supply-chain. An important example of this is the manner in which Principle 7 (Deliver Transparency and Auditability) is intended to be interpreted. In the Research and Invention phases of the technological life-cycle, compliance with Principle 7 requires understanding by inventors and innovators of the AI technology, and explicability to developers and users of AI-based artefacts and systems.

Table 2: Responsible A.I. Technologies, Artefacts, Systems and Applications
The 10 Principles

The following Principles apply to each entity responsible for each phase of AI research, invention, innovation, dissemination and application.

1. Evaluate Positive and Negative Impacts

AI offers prospects of considerable benefits and disbenefits. All entities involved in creating and applying AI have legal and moral obligations to assess the impacts, to demonstrate the benefits, to be proactive in relation to disbenefits, and to involve stakeholders in the process.

2. Complement Humans

Considerable public disquiet exists in relation to displacement of human workers by AI, and the replacement of human decision-making with inhumane machine decision-making.

3. Ensure Human Control

Considerable public disquiet exists in relation to the prospect of humans ceding power to machines.

4. Ensure Human Wellbeing and Safety

All entities involved in creating and applying AI have legal and moral obligations to provide safeguards for all human stakeholders who are at risk, whether as users of AI-based artefacts and systems, or as uses who are affected by them.

5. Ensure Consistency with Human Values and Human Rights

All entities involved in creating and applying AI have legal and moral obligations to address its negative impacts on the interests of individuals.

6. Deliver Transparency and Auditability

All entities have legal and moral obligations in relation to due process and procedural fairness. These obligations can only be fulfilled if all entities involved in creating and applying AI ensure that humanly-understandable explanations are available to the people who are affected by AI-based inferences, decisions and actions.

7. Embed Quality Assurance

All entities involved in creating and applying AI have legal and moral obligations in relation to the quality of business processes and products.

8. Exhibit Robustness and Resilience

All entities involved in creating and applying AI have legal and moral obligations to ensure safeguards are established and maintained, commensurate with the significance of the benefits, sensitivity and potential to cause harm to stakeholders.

9. Ensure Accountability for Legal and Moral Obligations

All entities involved in creating and applying AI have legal and moral obligations in relation to due process and procedural fairness. These obligations can only be fulfilled if each entity is discoverable, and each entity addresses problems as they arise.

10. Enforce, and Accept Enforcement of, Liabilities and Sanctions

All entities involved in creating and applying AI have legal and moral obligations in relation to due process and procedural fairness. These obligations can only be fulfilled if the entity implements internal problem-handling processes, and respects and complies with external problem-handling processes.

During the Innovation and Dissemination phases, the need is for understandability and manageability by developers and users of AI-based systems and applications, and explicability to affected stakeholders. In the Application phase, the emphasis shifts to understandability by affected stakeholders of inferences, decisions and actions arising from at least the AI elements within AI-based systems and applications.

The status of the proposed principles is important to appreciate. They are not expressions of law – although in some jurisdictions, and in some circumstances, some may of course be legal requirements. They are expressions of moral obligations; but no authority exists that can formally impose such obligations. All are contestable. They represent guidance to organisations involved in AI as to the expectations of courts, regulatory agencies, oversight agencies, competitors and stakeholders. They need to be taken into account as organisations undertake risk assessment and risk management.

In many circumstances, the Principles will be in conflict with other legal or moral obligations, and with various interests of various stakeholders. It can be argued, however, that each Principle creates an onus on each responsible entity. On that reading, each entity needs to ensure that the result of its endeavours is compliant, or that, to the extent that it is non-compliant, the organisation documents the factors that militate against compliance, documents the basis for its judgement that the importance of those factors outweighs that of the Principle, and diverts from the Principle only to the extent justified by those factors.

6. Application of the Principles

The Principles are intentionally framed and phrased in an abstract manner, in an endeavour to achieve general applicability. However, they lend themselves to customisation, and along multiple dimensions.

Firstly, as presaged in the earlier discussion of Principle 7, the language can be adapted to apply more precisely and clearly to each of the five phases of the AI supply chain. For example, the direct responsibilities of researchers and their employing institutions relate to the origination and potentialities of technologies, and the Principles can be phrased in ways familiar in that sector; whereas purveyors and users of AI artefacts and systems use the dialects of business and government, focus on application in the real world, and must directly account for impacts on individuals. However, each category of entities creates risks, and hence obligations exist along the entire AI supply chain. For any highly impactful technology, whether it is nuclear energy or AI, some responsibilities reach all the way to Phase 1, and to researchers.

Secondly, it will be beneficial for versions to be established that are specifically applicable to each form of AI. In the first instance, this can be done for the three currently mainstream forms discussed earlier – robotics, particularly remote-controlled and self-driving vehicles; cyborgisation; and AI/ML / neural-networking applications. For example, in a companion project, I have proposed 'Guidelines for Responsible Data Analytics' (Clarke 2018a). These are relevant to all forms of data analytics projects, including those that apply neural-networking approaches. Areas that they address include governance, expertise and compliance considerations,

multiple aspects of data acquisition and data quality, the suitability of both the data and the analytical techniques applied to it, and factors involved in the use of inferences drawn from the analysis.

The 'Principles for AI' are also capable of being further articulated into much more specific guidance in respect of sub-categories of AI technologies, artefacts, systems and applications. For example, sets could be spawned for each of vehicle collision-avoidance capabilities, specific prosthetic devices, and neural-network-based creditworthiness scoring schemes.

In addition to such primary uses of the Principles, they can be used as a basis for the *ex post facto* evaluation of existing technologies, artefacts, systems and applications. Further, independently-developed bodies of principles and guidelines can be compared with this set, in order to assess their comprehensiveness. That is of course not intended to imply that this set, based as it is on analysis of a mere 24 documents, is an uncontestable authority. Variances between sets may well lead to proposals for adaptation of this set, rather than to the deprecation of the alternative set. On the other hand, it is noteworthy that the average number of the 50 Principles detectable in the 24 documents is 8.6 (17%), and the 4 most comprehensive evidenced only 15, 16, 16 and 21 (30-42%).

Finally, the Principles might require modest re-casting in order to be readily applicable to what I proposed above as more appropriate conceptualisations of the field – complementary intelligence and intellectics.

Appendix 1: The 50 Principles

These are presented in a 2-page PDF format.

References

ACM (2017) 'Statement on Algorithmic Transparency and Accountability' Association for Computing Machinery, January 2017, at https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

Albus J. S. (1991) 'Outline for a theory of intelligence' IEEE Trans. Systems, Man and Cybernetics 21, 3 (1991) 473-509, at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.410.9719&rep=rep1&type=pdf>

Clarke R. (1989) 'Knowledge-Based Expert Systems: Risk Factors and Potentially Profitable Application Area', Xamax Consultancy Pty Ltd, January 1989, at <http://www.rogerclarke.com/SOS/KBTE.html>

Clarke R. (1993) 'Asimov's Laws of Robotics: Implications for Information Technology' in two parts, in IEEE Computer 26,12 (December 1993) 53-61, and 27,1 (January 1994) 57-66, at <http://www.rogerclarke.com/SOS/Asimov.html>

- Clarke R. (2005) 'Human-Artefact Hybridisation: Forms and Consequences' Proc. Ars Electronica 2005 Symposium on Hybrid - Living in Paradox, Linz, Austria, 2-3 September 2005, PrePrint at <http://www.rogerclarke.com/SOS/HAH0505.html>
- Clarke R. (2014) 'What Drones Inherit from Their Ancestors' Computer Law & Security Review 30, 3 (June 2014) 247-262, PrePrint at <http://www.rogerclarke.com/SOS/Drones-I.html>
- Clarke R. (2018a) 'Guidelines for the Responsible Application of Data Analytics' Computer Law & Security Review 34, 3 (May-Jun 2018) 467- 476, PrePrint at <http://www.rogerclarke.com/EC/GDA.html>
- Clarke R. (2018b) 'Principles for Responsible AI: Supporting Materials' Xamax Consultancy Pty Ltd, October 2018, at <http://www.rogerclarke.com/EC/PRAI-SM.html>
- CLA-EP (2016) 'Recommendations on Civil Law Rules on Robotics' Committee on Legal Affairs of the European Parliament, 31 May 2016, at <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN>
- FLI (2017) 'Asilomar AI Principles' Future of Life Institute, January 2017, at <https://futureoflife.org/ai-principles/?cn-reloaded=1>
- GEFA (2016) 'Position on Robotics and AI' The Greens / European Free Alliance Digital Working Group, November 2016, at <https://juliareda.eu/wp-content/uploads/2017/02/Green-Digital-Working-Group-Position-on-Robotics-and-Artificial-Intelligence-2016-11-22.pdf>
- IEEE (2017) 'Ethically Aligned Design', Version 2. IEEE, December 2017. at http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html
- ITIC (2017) 'AI Policy Principles' Information Technology Industry Council, undated but apparently of October 2017, at <https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf>
- McCarthy J. (2007) 'What is artificial intelligence?' Department of Computer Science, Stanford University, November 2007, at <http://www-formal.stanford.edu/jmc/whatisai/node1.html>
- Markoff J. (2013) 'In 1949, He Imagined an Age of Robots' The New York Times, 20 May 2013, at <http://www.nytimes.com/2013/05/21/science/mit-scholars-1949-essay-on-machine-age-is-found.html>
- Menzel P. & D'Alusio F. (2001) 'Robo sapiens' MIT Press, 2001
- Russell S.J. & Norvig P. (2009) 'Artificial Intelligence: A Modern Approach' Prentice Hall, 3rd edition, 2009
- Smith R. (2018). '5 core principles to keep AI ethical'. World Economic Forum, 19 Apr 2018, at <https://www.weforum.org/agenda/2018/04/keep-calm-and-make-ai-ethical/>
- Warwick K. (2014) 'The Cyborg Revolution' Nanoethics 8, 3 (Oct 2014) 263-273
-