Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/CLSR

Computer Law & Security Review

# Principles and business processes for responsible AI

*Roger Clarke*[a,b,c,*]

[a] *Xamax Consultancy Pty Ltd, Canberra, Australia*
[b] *Australian National University, Canberra, Australia*
[c] *University of N.S.W., Sydney, Australia*

ARTICLE INFO

*Keywords:*
AI artefacts
AI systems
AI ethics
AI risk assessment
AI risk management
AI Principles

ABSTRACT

The first article in this series examined why the world wants controls over Artificial Intelligence (AI). This second article discusses how an organisation can manage AI responsibly, in order to protect its own interests, but also those of its stakeholders and society as a whole. A limited amount of guidance is provided by ethical analysis. A much more effective approach is to apply adapted forms of the established techniques of risk assessment and risk management. Critically, risk assessment needs to be undertaken not only with the organisation's own interests in focus, but also from the perspectives of other stakeholders. To underpin this new form of business process, a set of Principles for Responsible AI is presented, consolidating proposals put forward by a diverse collection of 30 organisations.

## 1. Introduction

Proponents of Artificial Intelligence (AI) claim that it offers considerable promise, and some forms of it have indeed delivered value. On the other hand, the power inherent in AI naturally harbours substantial threats. The first round of threats afflicts the organisations that develop and deploy AI-based artefacts and systems. The second round of threats impacts upon the many categories of stakeholders that are involved in or are otherwise affected by the undertaking. Where stakeholders are aggrieved, and have sufficient power, their negative feedback affects the AI-originating organisations, particularly those that have direct associations with stakeholders, but also those further up the industry supply-chain.

This article adopts the position that it is in the interests of all organisations to avoid their stakeholders suffering harm, and thereby learning to distrust AI and oppose its use. For this to be achieved, organisations need to adopt responsible approaches to AI from the outset. This depends on the inculcation of appropriate culture within the organisation, and the establishment and operation of business processes whose purpose is to detect risk, and manage it.

This article first considers whether business ethics offers useful insights. A much more useful approach, however, is argued to be through risk assessment and risk management processes. These have traditionally had their focus on the interests of the organisation undertaking the study. Given the impactful nature of AI, this article proposes expansion of risk assessment and risk management in order to encompass not only the organisation's interests but also the perspectives of the various stakeholders. The final section draws on a diverse set of sources in order to propose a set of principles for the responsible application of AI, which are sufficiently specific to guide organisations' business processes.

[*] Corresponding author at: Xamax Consultancy Pty Ltd, 78 Sidaway St, Chapman ACT 2611, Australia.
*E-mail address:* roger.clarke@xamax.com.au

https://doi.org/10.1016/j.clsr.2019.04.007

## 2. Business ethics

Ethics is a branch of philosophy concerned with concepts of right and wrong conduct. Fieser (1995) and Pagallo (2016) distinguish 'meta-ethics', which is concerned with the language, origins, justifications and sources of ethics, from 'normative ethics', which formulates generic norms or standards, and 'applied ethics', which endeavours to operationalise norms in particular contexts.

From the viewpoint of instrumentalists in business and government, the field of ethics evidences several substantial deficiencies. The first is that there is no authority, or at least no uncontestable authority, for any particular formulation of norms, and hence every proposition is subject to debate. Further, as a form of philosophical endeavour, ethics embodies every complexity and contradiction that smart people can dream up. Moreover, few formulations by philosophers ever reach even close to operational guidance, and hence the sources enable prevarication and provide endless excuses for inaction. The inevitable result is that ethical discussions seldom have much influence on real-world behaviour. Ethics is an intellectually stimulating topic for the dinner-table, and graces *ex post facto* reviews of disasters. However, the notion of 'ethics by design' is even more empty than the 'privacy by design' meme. To an instrumentalist – who wants to get things done – ethics diversions are worse than a time-waster; they are a barrier to progress.

The periodically fashionable topic of 'business ethics' naturally inherits the vagueness of ethics generally (Donaldson and Dunfee, 1994; Joyner and Payne, 2002). Despite many years of discussion, the absence of any source of authoritative principles results in difficulties structuring concrete guidance for organisations in any of the many areas in which ethical issues are thought to arise. Far less does 'business ethics' assist in relation to complex and opaque digital technologies.

Clarke (2018b) consolidates a collection of applications of general ethical principles in technology-rich contexts – including bio-medicine, surveillance and information technology. Remarkably, none of them contains any explicit reference to identifying relevant stakeholders. However, a number of norms are apparent in multiple of the documents. These include demonstrated effectiveness and benefits, justification of disbenefits, mitigation of disbenefits, proportionality of negative impacts, supervision (including safeguards, controls and audit), and recourse (such as complaints and appeals channels, redress, sanctions, and enforcement powers and resources). These norms will be re-visited in the final section of this article.

The notion of Corporate Social Responsibility (CSR), which is sometimes extended to include an Environmental aspect, can be argued to have an ethical base (Carroll, 1999). CSR can extend beyond the direct interests of the organisation to include philanthropic contributions to individuals, community, society or the environment. In practice, however, its primary focus appears to be on the extraction of strategic advantage or public relations gains from organisations' required investments in regulatory compliance and their philanthropic activities (Porter and Kramer, 2006).

Although CSR and other aspects of the field of 'business ethics' have potential as a basis for establishing responsible approaches to AI, their effectiveness is constrained by the legal obligations imposed on company directors. Directors are required to act in the best interests of each company of which they are a director. Attention to broad ethical questions is extraneous to, and even in conflict with, that requirement, except where a business case indicates sufficient benefits to the organisation from taking a socially or environmentally responsible approach. In standard texts, for example, stakeholders are mentioned only as a factor in interpreting the director's duty to promote the success of the company (e.g. Keay, 2016). Guidance published by corporate regulators, directors' associations and major law firms generally omits mention of either ethics or social responsibility. Even developments in case law are at best only very slowly providing scope for directors to give meaningful consideration to stakeholders' interests (Marshall and Ramsay, 2009).

Stakeholders may benefit from organisations' compliance with regulatory requirements, and from their management of corporate reputation and of relationships with stakeholders. The categories of stakeholders that are most likely to have sufficient power to attract attention are large customers and suppliers, and employees. The scope might, however, extend to smaller customers, and to communities and economies on which the company has a degree of dependence. But this represents a slim basis on which to build a mechanism to achieve responsible AI.

The remainder of this article pursues more practical avenues. It assumes that organisations, when considering AI, begin by applying environmental scanning and marketing techniques in order to identify opportunities. They then use a business case approach to evaluate the strategic, market-share, revenue, cost and profit benefits that the opportunities appear to offer. The focus here is on how the downsides can be identified, evaluated and managed. Given the considerable investment that each organisation has in its culture, policies and procedures, it is desirable to align the approach so as to fit comfortably within or alongside the organisation's existing business processes.

## 3. Risk assessment and management

This section commences by briefly reviewing the conventional approach to the assessment and management of organisational risks. In order to address the risks that confront stakeholders, I contend that this framework must be extended. Beyond merely identifying stakeholders, the organisation needs to perform risk assessment and management from their perspectives. The second and third sub-sections accordingly consider stakeholder risk assessment and multi-stakeholder risk management.

### 3.1. *Organisational processes*

There are many sources of guidance in relation to organisational risk assessment and management. The techniques are particularly well-developed in the context of the security of IT assets and digital data, although the language and the approaches vary considerably among the many sources (most usefully: Firesmith, 2004; ISO, 2005; ISO, 2008; NIST, 2012; ENISA, 2016; ISM, 2017). For the present purpose, a model is adopted that is summarised in Appendix 1 of Clarke (2015). See Fig. 1.

The conventional approach to risk assessment and risk management is outlined in Table 1. Relevant organisational assets are identified, and an analysis is undertaken of the various forms of harm that could arise to those assets as a result of threats impinging on or actively exploiting vulnerabilities, and giving rise to threatening incidents. Existing safeguards are taken into account, in order to guide the development of a strategy and plan to refine and extend the safeguards. A degree of protection is sought that is judged to suitably balance modest actual costs against potentially much higher but contingent costs.

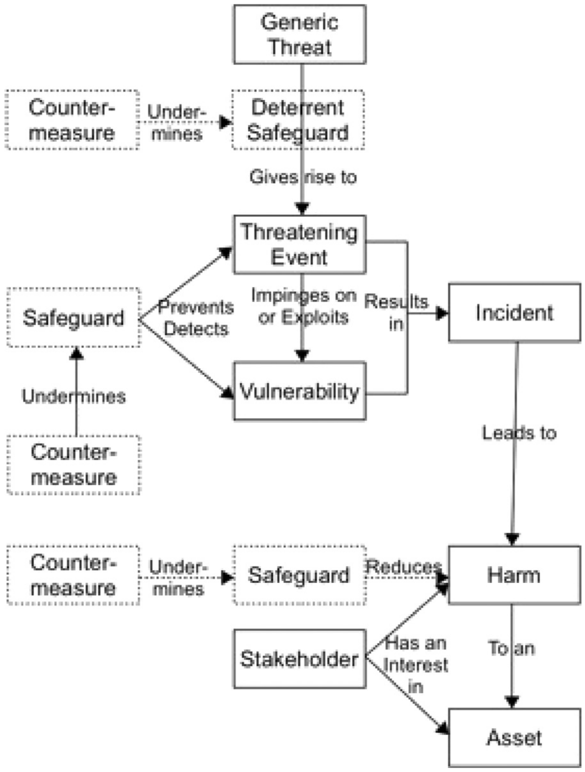

**Fig. 1 – The conventional risk model.**

**Table 1** – The risk assessment and risk management process.

**1. Analyse / Perform risk assessment**
1.1 Define the objectives and constraints
1.2 Identify the relevant stakeholders, assets, values and categories of harm
1.3 Analyse threats and vulnerabilities
1.4 Identify existing safeguards
1.5 Identify and prioritise the residual risks

**2. Design / Prepare for risk management**
2.1 Identify alternative safeguards
2.2 Evaluate the alternatives against the objectives and constraints
2.3 Select a design or adapt alternatives to achieve an acceptable design

**3. Do / Perform risk management**
3.1 Plan the implementation
3.2 Implement
3.3 Review the implementation

The initial, analysis phase provides the information needed for the strategy, design and planning processes whereby existing safeguards are adapted or replaced and new safeguards conceived and implemented. ISO standard 27005 (2008, pp. 20–24) discusses four options for what it refers to as 'risk treatment': risk modification, risk retention, risk avoidance and risk sharing. Table 2 presents a framework that in my experience is more understandable by practitioners and more readily usable as a basis for identifying possible safeguards.

Conventional approaches to risk assessment and management adopt the perspective of the organisation. The process identifies stakeholders, but their interests are reflected only to the extent that harm to them may result in material harm to the organisation. The focus of this article is on responsible behaviour by organisations in relation to the development and deployment of AI. As discussed in the previous article, AI is potentially highly impactful. Organisations accordingly need to

**Table 2 – Categories of risk management strategy.**

Proactive strategies
- **Avoidance**
  e.g. non-use of a risk-prone technology or procedure
- **Deterrence**
  e.g. signs, threats of dismissal, prosecutions, publicity for prosecutions, substantial fines, gaol-time
- **Prevention**
  e.g. quality software, physical and logical access control, designed and documented procedures, staff training, assigned responsibilities
- **Redundancy**
  e.g. duplicated equipment and communication paths, multiple, parallel evaluations with cross-checking of results

Reactive strategies
- **Detection**
  e.g. exception definitions, software-versioning, logging and time-stamping, log-analysis, exception reporting
- **Reduction/mitigation**
  e.g. suspension of processing when unexpected harm arises, pre-arranged contingent measures to compensate for harm
- **Recovery**
  e.g. designed and documented fallback procedures, staff training, assigned responsibilities, resource duplication including 'hot-sites' and 'warm-sites'
- **Insurance**
  e.g. maintenance contracts with suppliers, escrow of third party software, inspection of escrow deposits, policies with insurance companies

Non-reactive strategies
- **Tolerance/self-insurance**
  where assessment of the contingent costs concludes that they are bearable
- **Graceful degradation**
  e.g. a pre-funded compensation fund, combined with suspension or cancellation of processing when unexpected harm arises
- **Graceless degradation**
  e.g. preparedness to liquidate or disestablish the organisation when relatively very large unexpected harm arises

invest much more effort into understanding and addressing the risks faced by their stakeholders.

### 3.2. *Stakeholder risk assessment*

The term ‘stakeholders’ was coined, as a counterpoint to ‘shareholders’, in order to bring the interests of other parties into focus (Freeman and Reed, 1983). In IT contexts, users of information systems have long been recognised as stakeholders, commencing in the 1970s with employees. As inter-organisational and extra-organisational systems matured, IT services extended beyond organisations’ boundaries, and hence many suppliers and customers are now users as well.

The notion of ‘stakeholders’ is broader than just users, however. It comprises not only participants in information systems but also “any other individuals, groups or organizations whose actions can influence or be influenced by the development and use of the system whether directly or indirectly” (Pouloudi and Whitley, 1997, p. 3). The term ‘usees’ is usefully descriptive of such once-removed stakeholders (Clarke, 1992; Fischer-Huebner and Lindskog, 2001; Baumer, 2015). IT applications that have ‘usee stakeholders’ include credit bureau operations, shared databases about tenants and claimants on insurance policies, and intelligence systems operated by law enforcement agencies and private investigators. Further examples of ‘usees’ include employees’ dependants, local communities and the local physical environment, and, in the case of highly impactful IT, regional economies and natural ecologies.

Some AI projects may involve only a single major stakeholder group, such as employees. In many contexts, on the other hand, multiple stakeholders need to be recognised. For example, driverless vehicles affect not just passengers, but also occupants of other vehicles and pedestrians. The individuals in occupations whose existence is threatened by the new approach (e.g. taxi-, courier- and truck-drivers) expect to have a voice. So do their employers, and their unions, and those organisations may have sufficient influence to force their way into a place at the negotiation table. In the case of neural networking models, credit consumers, health insurance clients and welfare recipients may be so incensed by AI-derived discrimination against them that public pressure may force multi-stakeholder approaches onto lenders, health insurers and even welfare agencies – if only through politicians nervous about their electability. In the case of implanted medical devices, not only the patients, but also the various health care professions and health insurers have a stake in AI initiatives.

In the face of such complexities, how can an organisation effectively, but also efficiently, act responsibly in relation to AI projects?

### 3.3. *Multi-stakeholder risk management*

Conventional organisational risk assessment and risk management processes can be adapted in order to meet the need. My first proposition is that:

> *The responsible application of AI is only possible if stakeholder analysis is undertaken in order not only to identify the categories of entities that are or may be affected by the particular project, but also to gain insight into those entities’ needs and interests*

There are well-established techniques for conducting stakeholder analysis (Clarkson, 1995; Mitchell et al., 1997; Fletcher et al., 2003). There are also many commercially-published guidance documents. A natural tendency exists to focus on those entities that have sufficient market or institutional power to significantly affect the success of the project. On the other hand, in a world of social media and rapid and deep mood-swings, it is advisable to not overlook the nominally less powerful stakeholders. Where large numbers of individuals are involved (typically, employees, consumers and the general public), it will generally be practical to use representative and advocacy organisations as intermediaries, to speak on behalf of the categories or segments of individuals.

My second proposition is that:

> *Risk assessment processes that reflect the interests of stakeholders need to be broader than those commonly undertaken within organisations*

No term such as ‘public risk assessment’ appears to have become mainstream, but the concept of ‘impact assessment’ has. The earliest context was environmental impacts. Techniques more directly relevant to AI include the long-standing idea of ‘technology assessment’ (OTA, 1977), the little-developed field of social impact assessment (Becker and Vanclay, 2003), and the currently very active technique of ‘privacy impact assessment’ (Clarke, 2009; Wright and De Hert, 2012). For an example of impact assessment applied to the specific category of person-carrier robots, see Villaronga and Roig (2017).

My third proposition is that:

> *The responsible application of AI depends on risk assessment processes being conducted from the perspective of each stakeholder group, to complement that undertaken from the organisation’s perspective*

Such assessments could be conducted by the stakeholders independently, and fed into the organisation. However, the asymmetry of information, resources and power, and the degree of difference in world-views among stakeholder groups, may be so pronounced that the results of such independent activities may be difficult to assimilate and to integrate into the organisation’s ways of working.

Organisations may therefore find it advantageous to drive the studies and engage directly with the relevant parties. This enables the organisation to gain sufficiently deep understanding, and to reflect stakeholders’ needs in the project design criteria and features – and to do so without enormous cost to the organisation, and with the minimum harm to its own interests. Medical implants may provide good exemplars of multi-stakeholder risk assessment and management. These involve manufacturers undertaking carefully-designed pilot studies, with active participation by multiple health care professionals, patients, and patient advocacy organisations (ISO, 2011).

The risk assessment process outlined in Table 1 requires adaptation in order to reflect the broader set of interests under consideration. Table 3 depicts a possible approach.

This section has suggested customisation of existing, generic techniques in order to deal with the substantial impacts of AI-based systems. The following section considers more closely the specific requirements that need to be satisfied by multi-stakeholder risk management strategies, designs and plans.

## 4. Principles for responsible AI

The conduct of risk assessment from the perspectives of stakeholders depends on the analyst being attuned to those stakeholders' perspectives, interests and needs. This involves understanding and using a quite different vocabulary from that relevant to organisational risk assessment. In addition to those, already-substantial challenges, AI is a special case – arguably the most advanced, the most complex, the most mysterious, and the most threatening of all of the leaps forward that IT has gifted and imposed on the public during the 80 years since World War II stimulated invention, innovation and investment in computing.

Despite the frequent periods of public concern about IT, and despite the steady accretion of threats, the formulations of human rights that were negotiated during the post-War period have not yet been revised. So there is no ready-made body of norms, aspirational statements or expressions of moral rights on which risk analysts can depend.

There are, however, some sources of guidance. The bold claims of AI's proponents have generated counter-statements by academics and advocates for the interests of the public. Professional bodies in the IT arena have recently recognised that adaptation and articulation of their Codes of Ethics are long overdue, and have begun programs that are slowly moving beyond discussions and towards the expression of principles. Government bodies and supra-governmental associations have conducted studies. Corporations and industry associations have responded to these developments by uttering warm words of semi-commitment. Raw material exists.

This section presents a set of Principles for Responsible AI. Its purpose is to provide organisations and individual practitioners with guidance as to how they can fulfil their responsibilities in relation to AI technology and AI-based artefacts and systems. The Principles presented here are a natural fit to the needs of multi-stakeholder risk assessment and management, in particular to Activities 1.4–1.6 in Table 3, which involve the identification of assets, their value, and relevant threats, vulnerabilities and safeguards.

The following sub-section outlines the process whereby the Principles were drafted, and the kinds of sources that were used. The main body of the section explains the nature of the Principles, presents their abstract expression ('The 10 Themes'), and provides access to their more detailed expression ('The 50 Principles'). The final sub-section contains some meta-discussion about them.

### 4.1. *Sources and process*

The process of developing the set of Principles commenced with the postulation of Themes. This was based on prior reading in the fields of ethics in IT and AI, the analysis reported in the prior article in the present series, including the articulation of threats in s.4 of that article, and preliminary reading of many documents on possible safeguards.

Previously-published sets of principles were then catalogued and inspected. Diversity of perspective was actively sought, through searches in academic, professional and policy literatures. A total of 30 sources were identified and assessed. These were published variously by governmental organisations (9), corporations and industry associations (7), non-government organisations (6), academics (4), joint associations (2) and professional associations (2). Of the 30 source-documents, 8 are formulations of 'ethical principles and IT' (extracts and citations at Clarke, 2018b), and 22 focus on AI specifically (Clarke, 2019a).

**Table 3 – Multi-stakeholder risk assessment and risk management.**

| | | |
|---|---|---|
| **1. Analyse / Perform risk assessment** | | |
| 1.1 Define the objectives and constraints | | |
| 1.2 Identify the relevant stakeholders | | |
| **Organisation risk assessment** | **Stakeholder A risk assessment** | **Stakeholder B risk assessment** |
| 1.3 Objectives, constraints | 1.3 Objectives, constraints | 1.3 Objectives, constraints |
| 1.4 Assets, value, harms | 1.4 Assets, value, harms | 1.4 Assets, value, harms |
| 1.5 Threats, vulnerabilities | 1.5 Threats, vulnerabilities | 1.5 Threats, vulnerabilities |
| 1.6 Existing safeguards | 1.6 Existing safeguards | 1.6 Existing safeguards |
| 1.7 Residual risks | 1.7 Residual risks | 1.7 Residual risks |
| **2. Design / Prepare for risk management** | | |
| 2.1 Identify alternative safeguards | | |
| 2.2 Evaluate the alternatives against the objectives and constraints | | |
| 2.3 Select a design or adapt alternatives to achieve an acceptable design | | |
| **3. Do / Perform risk management** | | |
| 3.1 Plan the implementation | | |
| 3.2 Implement | | |
| 3.3 Review the implementation | | |

Of the significant documents that came to light, almost all were utilised. An exception was the inclusion of only one human rights document, the International Covenant on Civil and Political Rights (ICCPR, 1966), but the exclusion of other related treaties, regarding, for example, social and economic rights, and the rights of children, the disabled, refugees and older persons. An academic contribution that was omitted is Floridi et al. (2018). This mentions many possible principles, but its focus is processes whereby principles might be put into effect. Only sets that were available in the English language were used, resulting in a strong bias within the suite towards documents that originated in countries whose primary language(s) is or include English. The scope was nonetheless broad, in geographical terms, 11 from the USA, 9 from Europe, 6 from the Asia-Pacific, and 4 global in nature. In general, only the most recent version of documents available in March 2019 was used, except in the case of the European Commission, whose late 2018 draft and early 2019 final versions were both significant, and which evidence material differences.

Detailed propositions within each document were extracted, and allocated to themes, maintaining back-references to the sources. Where content from source-documents threw doubt on the structure or formulation of general Themes or specific Principles, the schema was adapted in order to sustain coherence and limit the extent to which duplications arise. The later documents were reasonably easily mapped onto the structure that had emerged from the earlier phases of analysis. Decreasing returns to scale suggested that the flex-point had been passed and that little further benefit would be gained from extending the set of source-documents.

Some items that appear in source-documents are not reflected in the Principles. For example, 'human dignity' and 'justice' are vague abstractions that need to be unpacked into more specific concepts. In addition, some proposals fall outside the scope of the present work. The items that have been excluded from the set are available as Supplementary Materials.

A similar approach to that described in this section was adopted by a research report, Zeng et al. (2019), which was published during the later stages of the preparation of the present article. Reviews of the contents of the Zeng article and the source-documents that those authors analysed did not suggest the need for any re-framing or re-phrasing of the 10 Themes or the 50 Principles.

In the previous article in this series, in s.4.5 and Table 2, distinctions were drawn among the successive phases of the supply-chain, which in turn produce AI technology, AI-based artefacts, AI-based systems, deployments of them, and applications of them. In each case, the relevant category of entity was identified that bears responsibility for negative impacts arising from AI. However, within the 30 sets of principles that were examined, only a few mentioned distinctions among entity-types, and in most cases it has to be interpolated which part of the supply-chain the document is intended to address. For example, 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 further observations about the nature of the Principles are offered in the final sub-section. The next section provides a brief preamble, presents the abstract set of 10 Themes, and provides access to the 50 Principles. A version of the 50 Principles that includes cross-references to the specific elements of each source-document is provided as Supplementary Materials.

### 4.2. *The themes and principles*

The status of the Principles enunciated here is important to appreciate. The purpose is to provide practical suggestions for organisations that are seeking to deal with AI responsibly, in particular by means of multi-stakeholder risk assessment and risk management. Each Theme and each Principle has been expressed in imperative mode, i.e. in the form of an instruction. This approach was adopted in order to convey that their purpose is to guide actions. They are not merely desirable characteristics, factors to be considered, or issues to be debated – although they can be used for those purposes as well.

The Principles represent guidance as to the expectations of stakeholders, but also of competitors, oversight agencies, regulatory agencies and courts. AI is a cluster of potentially powerful technologies and AI-based artefacts and systems may be applied within complex social contexts. The Principles are therefore not a recipe, but guidelines that require intelligent application. The Principles are not expressions of law – although in some jurisdictions, in some circumstances, some of them are legal requirements, and more may become so. They are expressions of moral obligations; but no authority exists that can impose such obligations. In addition, all of the Principles are contestable, and in different circumstances any of them may be in conflict with other legal or moral obligations, and with various other interests of various stakeholders. It is necessary to identify the need for trade-offs among interests and obligations, and to select appropriate balances.

In Table 4, the 10 Themes are stated, and brief explanations are provided whose purpose is orient the reader to the nature and intent of each Theme. This should ease the task of considering the detailed statements, which are in the Appendix to this article: The 50 Principles.

### 4.3. *Observations about the principles*

This sub-section contains a meta-discussion about several important aspects of the Principles. A first consideration is *whether the Principles address the threats* that were articulated in s.4 of the previous article in the series. In each case, this is achieved by means of a web of interlocking Principles. In particular, Autonomy is addressed in 2.1, 2.2, 3.1 and 3.2; Data Quality in 3.3, 6.2, 7.2 and 7.3; Process Quality in 1.3, 6.2, 7.6, 7.7, 7.8 and 7.9; Transparency in 2.1, 3.5, 6.1, 6.2, 6.3, 8.3 and 10.2; and Accountability in 1.6, 1.7, 1.8, 3.1, 3.2, 3.3, 9.1, 9.2, 10.1 and 10.2.

Some *commonalities* exist across some of the source documents. Overall, however, the main impression is of *sparse-*

**Table 4 – Responsible AI technologies, artefacts, systems and applications: the 10 Themes.**

The following apply to each entity responsible for each of the five phases of AI: research, invention, innovation, dissemination and application

**(1) Assess Positive and Negative Impacts and Implications**
AI offers prospects of considerable benefits and disbenefits. All entities involved in creating and applying AI have obligations to assess its short-term impacts and longer-term implications, to demonstrate the achievability of the postulated 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 the replacement of human decision-making by inhumane decision-making by AI-based artefacts and systems, and displacement of human workers by AI-based artefacts and systems.

**(3) Ensure Human Control**
Considerable public disquiet exists in relation to the prospect of humans being subject to obscure AI-based processes, and ceding power to AI-based artefacts and systems.

**(4) Ensure Human Safety and Wellbeing**
All entities involved in creating and applying AI have obligations to provide safeguards for all human stakeholders, whether as users of AI-based artefacts and systems, or as usees affected by them, and to contribute to human stakeholders' wellbeing.

**(5) Ensure Consistency with Human Values and Human Rights**
All entities involved in creating and applying AI have obligations to avoid, prevent and mitigate negative impacts on individuals, and to promote the interests of individuals.

**(6) Deliver Transparency and Auditability**
All entities have 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 affected by AI-based inferences, decisions and actions.

**(7) Embed Quality Assurance**
All entities involved in creating and applying AI have obligations in relation to the quality of business processes, products and outcomes.

**(8) Exhibit Robustness and Resilience**
All entities involved in creating and applying AI have obligations to ensure resistance to malfunctions (robustness) and recoverability when malfunctions occur (resilience), commensurate with the significance of the benefits, the data's sensitivity, and the potential for harm.

**(9) Ensure Accountability for Obligations**
All entities involved in creating and applying AI have obligations in relation to due process and procedural fairness. These obligations include the entity ensuring it is discoverable, and addressing problems as they arise.

**(10) Enforce, and Accept Enforcement of, Liabilities and Sanctions**
Each entity's obligations in relation to due process and procedural fairness include the implementation of systematic problem-handling processes, and respect for and compliance with external problem-handling processes.

*ness, with remarkably limited consensus*, particularly given that more than 60 years have passed since AI was first heralded. For example, only 1 document encompasses cyborgisation (GEFA, 2016); and only 2 documents refer to the precautionary principle (CLA-EP, 2016; GEFA, 2016).

An analysis was conducted of *the extent to which each source-document addresses the 50 Principles* that arose from the complete set. The analysis scored documents liberally, recognising them as delivering on a Principle if the idea was in some way evident, even if only some of the Principle was addressed, and irrespective of the strength of the prescription. Even then, the 30 documents reflected on average only 10 of the 50 Principles. Moreover, only 3 source-documents achieved moderately high scores – 46% for Marx (1998), 58% for EC (2018) and 74% for EC (2019). Apart from these three outliers, the mean was a very low 17%, range 8−34%. When assessed from the other direction, each of the 50 Principles was reflected in only 6 of the 30 documents (20%).

There was limited consensus even on quite fundamental requirements. 'Conduct impact assessment …' (Principle 1.4) was stipulated by only 11/30 documents. 'Ensure people's wellbeing ('beneficence')' (4.3) and 'Ensure that effective remedies exist …' (9.2) were each evident in only 14/30. Only one was visible in more than half, and two others were visible in precisely half of the sources:

- Ensure people's physical health and safety ('nonmaleficence') (4.1: 24/30)
- Be just/fair/impartial, treat individuals equally, and avoid unfair discrimination and bias … (5.1: 15/30)
- Ensure that data provenance, and the means whereby inferences are drawn, decisions are made, and actions are taken, are logged and can be reconstructed (6.2: 15/30)

Each of the sources naturally reflects the *express, implicit and subliminal purposes of the drafters and the organisations on whose behalf they were composed*. Among the sets of principles published by public interest organisations, the 4 academics achieved an average score of 23%, the 6 non-government organisations 23%, and the 9 government organisations 25%. On the other hand, the documents prepared by corporations, industry associations, joint associations and even professional associations tended to adopt the perspective of producer roles, with the interests of other stakeholders often relegated to a secondary consideration. The mean score for the 11 corporations, industry associations, joint associations of IT suppliers and users, and professional associations was barely half of that of the public interest organisations, at 13% (range 4−14% with two outliers at 22%, which were FLI (2017) and Sony (2019).

An example of the narrow approaches adopted by private sector organisations to date is that the joint-association Future Life Institute perceives the need for "constructive and healthy exchange between AI researchers and policy-makers", but not for any participation by stakeholders themselves (FLI, 2017 at 3). As a result, transparency is constrained to a small sub-set of circumstances (at 6), the degree of 'responsibility' of 'designers and builders' is limited to those roles being mere '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, emphasis added), and "strict safety and control measures" are limited to a small subset of AI systems (at 22).

The authors of ITIC (2017) consider that many responsibilities lie elsewhere, and responsibilities are assigned 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 regards decision opaqueness as being acceptable. 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 (HOL, 2018) might have been expected to adopt a societal or multi-stakeholder approach, yet it appears to have adopted the perspective of the AI industry.

Some of the Principles require *different interpretation in each phase of the AI supply-chain*. An important example of this is the manner in which Theme 7 – Deliver Transparency and Auditability – is intended to be applied. In the Research and Invention phases of the technological life-cycle, compliance with Theme 7 requires understanding by inventors and innovators of the AI technology, and explicability to developers and users of AI-based artefacts and systems. 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.

Some of the source-documents primarily address just one form of AI, such as robotics or machine-learning. The consolidated Principles, however, were *framed and phrased in a reasonably general manner*, in an endeavour to achieve applicability to at least the AI technologies discussed in the first article in the series – robotics, particularly remote-controlled and self-driving vehicles; cyborgs who incorporate computational capabilities; rule-based expert systems; and AI/ML/neural-networking applications. More broadly, the intention is that the Principles be applicable to what I proposed in the first article in the series as more appropriate conceptualisations of the field – Complementary Artefact Intelligence, and Intellectics.

The Principles are *capable of being further articulated* into much more specific guidance in respect of each particular AI technology. For example, in a companion project, I have proposed 'Guidelines for Responsible Data Analytics' (Clarke, 2018a). These provide more detailed guidance for the conduct of all forms of data analytics projects, including those that apply AI/ML/neural-networking approaches. The 50 Principles can also be customised to the needs of organisations operating in particular roles within the AI industry supply-chain, identified in Table 2 of the previous article as Researchers, R&D Engineers, Developers, Purveyors, and User Organisations and Individuals.

## 5. Conclusions

AI technologies emerge from research laboratories offering potential, but harbouring considerable threats to both organisations and their stakeholders. The first article in this series examined why there is public demand for controls over AI, and argued that organisations need to adopt, and to demonstrate publicly that they have adopted, responsible approaches to AI.

This article has formulated guidance for organisations, in both the private and public sectors, whereby they can evaluate the appropriateness of AI technologies to their own operations. Ethical analysis does not deliver what organisations need. Adapted forms of risk assessment and risk management processes have been proposed to meet the requirements. A set of 50 Principles for Responsible AI has been presented, derived from 30 documents from diverse sources.

The Principles are capable of being applied by executives, managers and professionals, as they stand. They can also be used as a basis for customising Principles for specific AI technologies, for particular categories of AI artefacts, systems and applications, and for organisations operating in various parts of the AI industry supply-chain. Once products have been developed and deployed, the Principles provide a reference-point for reviews by internal and external auditors.

The Principles are appropriate as a template for the expression of organisational procedures, or industry codes, or indeed legislation; or as a measuring stick for the evaluation of proposals for such documents. As indicated above, the source-document that is far more comprehensive than any other published to date is the European Commission's 'Ethics Guidelines for Trustworthy AI' (EC, 2019), which was published shortly before finalisation of this article. As this article went to press, the OECD published 'Recommendations on AI' (OECD 2019). An evaluation of the OECD's Principles against the 50 Principles presented in this article gave rise to a score of only 40% (Clarke 2019b). The European Commission's document therefore continues to stand alone as official guidance with moderately good coverage. It is noteworthy, however, that the EC's, 2019 Guidelines still only achieve a score of 74%. The 50 Principles, being a consolidation of a significant population of source-documents, represents the most comprehensive available guidance.

This second article in the series has as its focus what organisations can do in order to develop and apply AI responsibly. The third article considers how policy-makers can structure a regulatory regime to ensure that, whether or not individual organisations approach AI in a responsible manner, important public interests can nonetheless be protected.

## Conflict of interest

None.

## Acknowledgement

This paper has benefited from feedback from multiple colleagues, and particularly Peter Leonard of Data Synergies, Prof. Graham Greenleaf and Kayleen Manwaring of UNSW, and Prof. Tom Gedeon of ANU. The comments of an anonymous referee were also helpful in ensuring clarification of key elements of the argument.

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.clsr.2019.04.007.

## Appendix 1. Responsible AI technologies, artefacts, systems and applications: the 50 principles

A PDF version of this Appendix is available at: http://www.rogerclarke.com/EC/AIP-App1.pdf.

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

1. **Assess positive and negative impacts and implications**

1.1 Conceive and design only after ensuring adequate understanding of purposes and contexts
1.2 Justify objectives
1.3 Demonstrate the achievability of postulated benefits
1.4 Conduct impact assessment, including risk assessment from all stakeholders' perspectives
1.5 Publish sufficient information to stakeholders to enable them to conduct their own assessments
1.6 Conduct consultation with stakeholders and enable their participation in design
1.7 Reflect stakeholders' justified concerns in the design
1.8 Justify negative impacts on individuals ('proportionality')
1.9 Consider alternative, less harmful ways of achieving the same objectives

2. **Complement humans**

2.1 Design as an aid, for augmentation, collaboration and inter-operability
2.2 Avoid design for replacement of people by independent artefacts or systems, except in circumstances in which those artefacts or systems are demonstrably more capable than people, and even then ensuring that the result is complementary to human capabilities

3. **Ensure human control**

3.1 Ensure human control over AI-based technology, artefacts and systems
3.2 In particular, ensure human control over autonomous behaviour of AI-based technology, artefacts and systems
3.3 Respect people's expectations in relation to personal data protections, including:
- their awareness of data-usage
- their consent
- data minimization
- public visibility and design consultation and participation
- the relationship between data-usage and the data's original purpose

3.4 Respect each person's autonomy, freedom of choice and right to self-determination
3.5 Ensure human review of inferences and decisions before action is taken
3.6 Avoid deception of humans
3.7 Avoid services being conditional on the acceptance of AI-based artefacts and systems

4. **Ensure human safety and wellbeing**

4.1 Ensure people's physical health and safety ('nonmaleficence')
4.2 Ensure people's psychological safety, by avoiding negative effects on their mental health, emotional state, inclusion in society, worth, and standing in comparison with other people
4.3 Contribute to people's wellbeing ('beneficence')
4.4 Implement safeguards to avoid, prevent and mitigate negative impacts and implications
4.5 Avoid violation of trust
4.6 Avoid the manipulation of vulnerable people, e.g. by taking advantage of individuals' tendencies to addictions such as gambling, and to letting pleasure overrule rationality

5. **Ensure consistency with human values and human rights**

5.1 Be just/fair/impartial, treat individuals equally, and avoid unfair discrimination and bias, not only where they are illegal, but also where they are materially inconsistent with public expectations
5.2 Ensure compliance with human rights laws
5.3 Avoid restrictions on, and promote, people's freedom of movement
5.4 Avoid interference with, and promote privacy, family, home or reputation
5.5 Avoid interference with, and promote, the rights of freedom of information, opinion and expression, of freedom of assembly, of freedom of association, of freedom to participate in public affairs, and of freedom to access public services
5.6 Where interference with human values or human rights is outweighed by other factors, ensure that the interference is no greater than is justified ('harm minimisation')

6. **Deliver transparency and auditability**

6.1 Ensure that the fact that a process is AI-based is transparent to all stakeholders
6.2 Ensure that data provenance, and the means whereby inferences are drawn from it, decisions are made, and actions are taken, are logged and can be reconstructed
6.3 Ensure that people are aware of inferences, decisions and actions that affect them, and have access to humanly-understandable explanations of how they came about

7. **Embed quality assurance**

7.1 Ensure effective, efficient and adaptive performance of intended functions
7.2 Ensure data quality and data relevance
7.3 Justify the use of data, commensurate with each data-item's sensitivity
7.4 Ensure security safeguards against inappropriate data access, modification and deletion, commensurate with its sensitivity
7.5 Deal fairly with people ('faithfulness', 'fidelity')
7.6 Ensure that inferences are not drawn from data using invalid or unvalidated techniques
7.7 Test result validity, and address the problems that are detected
7.8 Impose controls in order to ensure that the safeguards are in place and effective
7.9 Conduct audits of safeguards and controls

8. **Exhibit robustness and resilienc**e

8.1 Deliver and sustain appropriate security safeguards against the risk of compromise of intended functions arising from both passive threats and active attacks, commensurate with the significance of the benefits and the potential to cause harm
8.2 Deliver and sustain appropriate security safeguards against the risk of inappropriate data access, modification and deletion, arising from both passive threats and active attacks, commensurate with the data's sensitivity
8.3 Conduct audits of the justification, the proportionality, the transparency, and the harm avoidance, prevention and mitigation measures and controls
8.4 Ensure resilience, in the sense of prompt and effective recovery from incidents

9. **Ensure accountability for obligations**

9.1 Ensure that the responsible entity is apparent or can be readily discovered by any party
9.2 Ensure that effective remedies exist, in the form of complaints processes, appeals processes, and redress where harmful errors have occurred

10. **Enforce, and accept enforcement of, liabilities and sanctions**

10.1 Ensure that complaints, appeals and redress processes operate effectively
10.2 Comply with external complaints, appeals and redress processes and outcomes, including, in particular, provision of timely, accurate and complete information relevant to cases.

## REFERENCES

ACM. Statement on algorithmic transparency and accountability. Association for Computing Machinery; 2017 January 2017, at https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

Baumer EPS. Usees. Proceedings of the 33rd annual ACM conference on human factors in computing systems (CHI'15), 2015.

Becker H. & Vanclay F. 'The international handbook of social impact assessment' Cheltenham: Edward Elgar, 2003.

Carroll AB. Corporate social responsibility: evolution of a definitional construct. Bus Soc 1999;38(3):268–95.

CLA-EP. Recommendations on civil law rules on robotics. Committee on Legal Affairs of the European Parliament; 2016 31 May 2016, at http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN.

Clarke R. Extra-organisational systems: a challenge to the software engineering paradigm. Proceedings of the IFIP world congress, 1992. http://www.rogerclarke.com/SOS/PaperExtraOrgSys.html.

Clarke R. Privacy impact assessment: its origins and development. Comput Law Secur Rev 2009;25(2):123–35, PrePrint at http://www.rogerclarke.com/DV/PIAHist-08.html.

Clarke R. The prospects of easier security for SMEs and consumers. Comput Law Secur Rev 2015;31(4):538–52. PrePrint at http://www.rogerclarke.com/EC/SSACS.html.

Clarke R. Guidelines for the responsible application of data analytics. Comput Law Secur Rev 2018a;34(3):467–76. PrePrint at http://www.rogerclarke.com/EC/GDA.html.

Clarke R. Ethical principles and information technology. Xamax Consultancy Pty Ltd; 2018b rev. September 2018, at http://www.rogerclarke.com/EC/GAIE.html.

Clarke R. Principles for AI: a 2017–18 SourceBook. Xamax Consultancy Pty Ltd; 2019a rev. at http://www.rogerclarke.com/EC/GAIP.html.

Clarke R. The OECD's AI Guidelines of 22 May 2019: Evaluation against a Consolidated Set of 50 Principles. Xamax Consultancy Pty Ltd; 2019b 26 May 2019, at http://www.rogerclarke.com/EC/AI-OECD-Eval.html.

Clarkson MBE. A stakeholder framework for analyzing and evaluating corporate social performance. Acad Manag Rev 1995;20(1):92–117, at https://www.researchgate.net/profile/Mei_Peng_Low/post/Whats_corporate_social_performance_related_to_CSR/attachment/59d6567879197b80779ad3f2/AS%3A530408064417792%401503470545971/download/A_Stakeholder_Framework_for_Analyzing+CSP.pdf.

Donaldson T, Dunfee TW. Toward a unified conception of business ethics: integrative social contracts theory. Acad Manag Rev 1994;19(2):252–84.

EC. Statement on artificial intelligence, robotics and 'autonomous' systems. European Commission; 2018 European Group on Ethics in Science and New Technologies March 2018, at http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

EC. Ethics guidelines for trustworthy AI. European Commission; 2019 High-Level Expert Group on Artificial Intelligence, at https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477.

ENISA. Risk management: implementation principles and inventories for risk management/risk assessment methods and tools. European Union Agency for Network and Information Security; 2016 June 2016, at https://www.enisa.europa.eu/publications/risk-management-principles-and-inventories-for-risk-management-risk-assessment-methods-and-tools.

Fieser J. Ethics. Internet Encyclopedia of Philosophy; 1995, at https://www.iep.utm.edu/ethics/.

Firesmith D. Specifying reusable security requirements. J Object Technol 2004;3(1):61–75. Jan–Feb 2004 at http://www.jot.fm/issues/issue_2004_01/column6.

Fischer-Huebner S, Lindskog H. Teaching privacy-enhancing technologies. Proceedings of the IFIP WG 11.8 2nd world conference on information security education; 2001. p. 2001, at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.24.3950&rep=rep1&type=pdf.

Fletcher A, Guthrie J, Steane P, Roos G, Pike S. Mapping stakeholder perceptions for a third sector organization. J Intellect Cap 2003;4(4):505–27 2003.

FLI. Asilomar AI principles. Future of Life Institute; 2017 January 2017, at https://futureoflife.org/ai-principles/.

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E. People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach 2018;28(2018):689–707.

Freeman RE, Reed DL. Stockholders and stakeholders: a new perspective on corporate governance. Calif Manag Rev 1983;25(3):88–106, at https://www.researchgate.net/profile/R_Freeman/publication/238325277_Stockholders_and_Stakeholders_A_New_Perspective_on_Corporate_Governance/links/5893a4b2a6fdcc45530c2ee7/Stockholders-and-Stakeholders-A-New-Perspective-on-Corporate-Governance.pdf.

GEFA. Position on robotics and AI. The Greens/European Free Alliance; 2016 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.

HOL. AI in the UK: ready, willing and able?. House of Lords; 2018 Select Committee on Artificial Intelligence April 2018, at https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

ICCPR. International covenant on civil and political rights, 1966. United Nations; 1966, at http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx.

IEEE. Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems (A/IS). IEEE; 2017 Version 2, December 2017, at http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

ISM. Information security manual. Australian Signals Directorate; 2017 November 2017, at https://acsc.gov.au/infosec/ism/index.htm.

ISO. Information technology – code of practice for information security management, 27002. International Standards Organisation, ISO/IEC; 2005.

ISO. Information technology – security techniques – information security risk management, 27005. ISO/IEC; 2008.

ISO. Clinical investigation of medical devices for human subjects – good clinical practice, 14155. ISO; 2011.

ITIC. AI policy principles. Information Technology Industry Council; 2017 undated but apparently of October 2017, at https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf.

Joyner BE, Payne D. Evolution and implementation: a study of values, business ethics and corporate social responsibility. J Bus Ethics 2002;41(4):297–311.

Keay A. Directors' duties. Lexis-Nexis 2016.

Marshall S, Ramsay I. Shareholders and directors' duties: law, theory and evidence. Melbourne Law School, University of Melbourne; 2009 Legal Studies Research Paper No. 411 June 2009, at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1402143.

Marx GR. An ethics for the new surveillance, 14. The Information Society; 1998. p. 171–85.

Mitchell RK, Agle BR, Wood DJ. Toward a theory of stakeholder identification and salience: defining the principle of who and what really counts. Acad Manag Rev 1997;22(4):853–86.

NIST. Guide for conducting risk assessments 2012. Special Publication SP 800-30 Rev. 1, September 2012, at http://csrc.nist.gov/publications/nistpubs/800-30-rev1/sp800_30_r1.pdf.

OECD. Recommendation of the Council on Artificial Intelligence 2019. Organisation for Economic Co-operation and Development, 22 May 2019, at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

OTA. Technology assessment in business and government 1977. NTIS order #PB-273164′, January 1977, at http://www.princeton.edu/~ota/disk3/1977/7711_n.html.

Pagallo U. Even angels need the rules: AI, roboethics, and the law. Proceedings of the ECAI, 2016.

Porter ME, Kramer MR. The link between competitive advantage and corporate social responsibility. Harvard Bus Rev 2006;84(12):78–92.

Pouloudi A, Whitley EA. Stakeholder identification in inter-organizational systems: gaining insights for drug use management systems. Eur J Inf Syst 1997;6(1):1–14, at http://eprints.lse.ac.uk/27187/1/__lse.ac.uk_storage_LIBRARY_Secondary_libfile_shared_repository_Content_Whitley_Stakeholder%20identification_Whitley_Stakeholder%20identification_2015.pdf.

Sony. Sony group AI ethics guidelines. Sony; 2019. 1 Mar 2019, at https://www.sony.net/SonyInfo/csr_report/humanrights/hkrfmg0000007rtj-att/AI_Engagement_within_Sony_Group.pdf.

Villaronga EF, Roig A. European regulatory framework for person carrier robots. Comput Law Secur Rev 2017;33(4):502–20.

Wright D, De Hert P. Privacy impact assessments. Springer; 2012.

Zeng Y, Lu E, Huangfu C. Linking artificial intelligence principles. Proceedings of the AAAI workshop on artificial intelligence safety (AAAI-Safe AI), 2019. at https://arxiv.org/abs/1812.04814