Available online at www.sciencedirect.com

ScienceDirect

www.compseconline.com/publications/prodclaw.htm

Computer Law & Security Review

Guidelines for the responsible application of data analytics

Roger Clarke [a,b,c,*]

[a] *Xamax Consultancy Pty Ltd, Canberra, Australia*
[b] *University of NSW Law, Sydney, Australia*
[c] *Research School of Computer Science, Australian National University, Canberra, Australia*

Keywords:
Big data
Data science
Data quality
Decision quality
Regulation

ABSTRACT

The vague but vogue notion of 'big data' is enjoying a prolonged honeymoon. Well-funded, ambitious projects are reaching fruition, and inferences are being drawn from inadequate data processed by inadequately understood and often inappropriate data analytic techniques. As decisions are made and actions taken on the basis of those inferences, harm will arise to external stakeholders, and, over time, to internal stakeholders as well. A set of Guidelines is presented, whose purpose is to intercept ill-advised uses of data and analytical tools, prevent harm to important values, and assist organisations to extract the achievable benefits from data, rather than dreaming dangerous dreams.

1. Introduction

Previous enthusiasms for management science, decision support systems, data warehousing and data mining have been rejuvenated. Fervour for big data, big data analytics and data science has been kindled, and is being sustained, by high-pressure technology salesmen. Like all such fads, there is a kernel of truth, but also a large penumbra of misunderstanding and misrepresentation, and hence considerable risk of disappointment, and worse.

A few documents have been published that purport to provide some advice on how to avoid harm arising from the practice of these techniques. Within the specialist big data analytics literature, the large majority of articles focus on techniques and applications, with impacts and implications relegated to a few comments at the end of the paper rather than even being embedded within the analysis, let alone a driving factor in the design. But see Agrawal et al. (2011), Saha and Srivastava (2014), Jagadish et al. (2014), Cai and Zhu (2015) and Haryadi et al. (2016), and particularly Merino et al. (2016).

Outside academe, most publications that offer advice appear to be motivated not by the avoidance of harm to affected values, but rather the protection of the interests of organisations conducting analyses and using the results. Examples of such documents in the public sector include DoFD (2015) – subsequently withdrawn, and UKCO (2016). Nothing resembling guidelines appears to have been published to date by the relevant US agencies, but see NIST (2015) and GAO (2016).

Some professional codes and statements are relevant, such as UNSD (1985), DSA (2016), ASA (2016) and ACM (2017). Examples also exist in the academic research arena, e.g. Rivers and Lewis (2014), Müller et al. (2016) and Zook et al. (2017). However, reflecting the dependence of the data professions on the freedom to ply their trade, such documents are oriented towards facilitation, with the protection of stakeholders commonly treated as a constraint rather than as an objective.

* *Corresponding author.* Xamax Consultancy Pty Ltd, 78 Sidaway St, Chapman ACT 2611 Canberra, Australia.
E-mail address: Roger.Clarke@xamax.com.au (R. Clarke).
https://doi.org/10.1016/j.clsr.2017.11.002
0267-3649/

Documents have begun to emerge from government agencies that perform regulatory rather than stimulatory functions. See, for example, a preliminary statement issued by Data Protection Commissioners (WP29, 2014), a consultation draft from the Australian Privacy Commissioner (OAIC, 2016), and a document issued by the Council of Europe Convention 108 group (CoE 2017). These are, however, unambitious and diffuse, reflecting the narrow statutory limitations of such organisations to the protection of personal data. For a more substantial discussion paper, see ICO (2017).

It is vital that guidance be provided for at least those practitioners who are concerned about the implications of their work. In addition, a reference-point is needed as a basis for evaluating the adequacy of organisational practices, of the codes and statements of industry and professional bodies, of recommendations published by regulatory agencies, and of the provisions of laws and statutory codes. This paper's purpose is to offer such a reference-point, expressed as guidelines for practitioners who are seeking to act responsibly in their application of analytics to big data collections.

This paper draws heavily on previous research reported in Wigan and Clarke (2013), Clarke (2016a, 2016b), Raab and Clarke (2016) and Clarke (2017b). It also reflects literature critical of various aspects of the big data movement, notably Bollier (2010), Boyd and Crawford (2011), Lazer et al. (2014), Metcalf and Crawford (2016), King and Forder (2016) and Mittelstadt et al. (2016). It first provides a brief overview of the field, sufficient to provide background for the remainder of the paper. It then presents a set of Guidelines whose intentions are to filter out inappropriate applications of data analytics, and provide a basis for recourse by aggrieved parties against organisations whose malbehaviour or misbehaviour results in harm. An outline is provided of various possible applications of the Guidelines.

2. Background

The 'big data' movement is largely a marketing phenomenon. Much of the academic literature has been cavalier in its adoption and reticulation of vague assertions by salespeople. As a result, definitions of sufficient clarity to assist in analysis are in short supply. This author adopts the approach of treating as 'big data' any collection that is sufficiently large that someone is interested in applying sophisticated analytical techniques to it. However, it is important to distinguish among several categories:

- a single large data collection; and
- a consolidation of two or more data collections, which may be achieved through:
  - merger into a single physical data collection; or
  - interlinkage into a single virtual data collection

The term 'big data analytics' is distinguishable from its predecessor 'data mining' primarily on the basis of the decade in which it is used. It is subject to marketing hype to almost the same extent as 'big data'. So all-inclusive are its usages that a reasonable working definition is:

> *Big data analytics encompasses all processes applied to big data that may enable inferences to be drawn from it.*

The term 'data scientist' emerged two decades ago as an upbeat alternative to 'statistician' (Press, 2013). Its focus is on analytic techniques, whereas the more recent big data movement commenced with its focus on data. The term 'data science' has been increasingly co-opted by the computer science discipline and business communities in order to provide greater respectability to big data practices. Although computer science has developed some additional techniques, a primary focus has been the scalability of computational processes to cope with large volumes of disparate data. It may be that the re-capture of the field by the statistics discipline will bring with it a recovery of high standards of professionalism and responsibility – which, this paper argues, are sorely needed. In this paper, however, the still-current term 'big data analytics' is used.

Where data is not in a suitable form for application of any particular data analytic technique, modifications may be made to it in an attempt to address the data's deficiencies. This was for many years referred to as 'data scrubbing', but it has become more popular among proponents of data analytics to use the misleading terms 'data cleaning' and 'data cleansing' (e.g. Rahm and Do, 2000, Müller and Freytag, 2003). These terms imply that the scrubbing process reliably achieves its aim of delivering a high-quality data collection. Whether that is actually so is highly contestable, and is seldom demonstrated through testing against the real world that the modified data purports to represent. There are many challenging aspects of data quality. What should be done where data-items that are important to the analysis are empty ('null')? And what should be done where they contain values that are invalid according to the item's definition, or have been the subject of varying definitions over the period during which the data-set has been collected? Another term that has come into currency is 'data wrangling' (Kandel et al., 2011). Although the term is honest and descriptive, and the authors adopt a systematic approach to the major challenge of missing data, their processes for 'correcting erroneous values' are merely computationally-based 'transforms', neither sourced from nor checked against the real world. The implication that data is 'clean' or 'cleansed' is commonly an over-claim, and hence such terms should be avoided in favour of the frank and usefully descriptive term 'data scrubbing'.

Where data is consolidated from two or more data collections, some mechanism is needed to determine which records in each collection are appropriately merged or linked. In some circumstances there may be a common data-item in each collection that enables associations between records to be reliably postulated. In many cases, a combination of data-items (e.g., in the case of people, the set of first and last name, date-of-birth and postcode) may be regarded as representing the equivalent of a common identifier. This process has long been referred to as computer or data matching (Clarke, 1994). Other approaches can be adopted, but generally with even higher incidences of false-positives (matches that are made but that are incorrect) and false-negatives (matches that could have been made but were not). A further issue is the extent to which a consolidated collection should contain all entries or only those for which a match has (or has not) been found. This decision may have a significant

impact on the usability of the collection, and on the quality of inferences drawn from it.

Significantly, descriptions of big data analytics processes seldom make any provision for a pre-assessment of the nature and quality of the data that is to be processed. See, for example, Jagadish (2015) and Cao (2017). Proponents of big data analytics are prone to make claims akin to 'big trumps good', and that data quality is irrelevant if enough data is available. Circumstances exist in which such claims may be reasonable; but for most purposes they are not (Bollier, 2010; Boyd and Crawford, 2011; Clarke, 2016a), and data quality is an important consideration. McFarland and McFarland (2015) argue that 'precisely inaccurate' results arise from the 'biased samples' that are an inherent feature of big data.

A structured framework for assessing data quality is presented in Table 1. It draws on a range of sources, importantly Huh et al. (1990), Wang and Strong (1996), Müller and Freytag (2003) and Piprani and Ernst (2008). See also Hazen et al. (2014). Each of the factors in the first group can be assessed at the time of data acquisition and subsequently, whereas those in the second group, distinguished as 'information quality' factors, can only be judged at the time of use.

Underlying these factors are features of data that are often overlooked, but that become very important in the 'big data' context of data expropriation, re-purposing and merger. At the heart of the problem is the materially misleading presumption that data is 'captured'. That which pre-exists the act of data collection comprises real-world phenomena, not data that is available for 'capture'. Each item of data is created, by a process performed by a human or an artefact that senses the world and records a symbol that is intended to represent some aspect of the phenomena that is judged to be relevant. The choice of phenomena and of their attributes, and the processes for creating data to represent them, are designed and implemented by or on behalf of some entity that has some purpose in mind. The effort invested in data quality assurance at the time that it is created reflects the characteristics of the human or artefact that creates it, the process whereby it is created, the purpose of the data, the value of the data and of data quality to the relevant entity, and the available resources. Hence the relationship between the data-item and the real world phenomenon that it purports to represent is not infrequently tenuous, and is subject to limitations of definition, observation, measurement, accuracy, precision and cost.

The conduct of data analytics also depends heavily on the meanings imputed to data-items. Uncertainties arise even within a single data collection. Where a consolidated collection is being analysed, inferences may be drawn based on relationships among data-items that originated from different sources. The reasonableness of the inferences is heavily dependent not only on the quality and meaning of each item, but also on the degree of compatibility among their quality profiles and meanings.

A further serious concern is the propensity for proponents of big data to rely on correlations, without any context resembling a causative model. This even extends to championing the death of theory (Anderson, 2008; Mayer-Schonberger and Cukier, 2013). Further, it is all too common for proponents of big data analytics to interpret correlations as somehow being predictive, and then apply them as if they were prescriptive.

When big data analytics techniques are discussed, the notion of Artificial Intelligence (AI) is frequently invoked. This is a catch-all term that has been used since the mid-1950s. Various strands have had spurts of achievement, particularly in the pattern-matching field, but successes have been interspersed

Table 1 – Quality factors.

Data Quality Factors	
(assessable at the time of creation and subsequently)	
D1	**Syntactic Validity** Conformance of the data with the domain on which the data-item is defined
D2	**Appropriate (Id)entity Association** A high level of confidence that the data is associated with the particular real-world identity or entity whose attribute(s) it is intended to represent
D3	**Appropriate Attribute Association** The absence of ambiguity about which real-world attribute(s) the data is intended to represent
D4	**Appropriate Attribute Signification** The absence of ambiguity about the particular state of the particular real-world attribute(s) that the data is intended to represent
D5	**Accuracy** A high degree of correspondence of the data with the real-world phenomenon that it is intended to represent, typically measured by a confidence interval, such as '±1 degree Celsius'
D6	**Precision** The level of detail at which the data is captured, reflecting the domain on which valid contents for that data-item are defined, such as 'whole numbers of degrees Celsius'
D7	**Temporal Applicability** The absence of ambiguity about the date and time when, or the period of time during which, the data represents or represented a particular real-world phenomenon. This is important in the case of volatile data-items such as total rainfall for the last 12 months, marital status, fitness for work, age, and the period during which an income-figure was earned or a licence was applicable
Information Quality Factors	
(assessable only at the time of use)	
I1	**Theoretical Relevance** A demonstrable capability of the data-item to make a difference to the inferencing process in which the data is to be used
I2	**Practical Relevance** A demonstrable capability of the data-item's content to make a difference to the inferencing process in which the data is to be used
I3	**Currency** The absence of a material lag between a real-world occurrence and the recording of the corresponding data
I4	**Completeness** The availability of sufficient contextual information that the data is not liable to be misinterpreted
I5	**Controls** The application of business processes that ensure that the data quality and information quality factors have been considered prior to the data's use
I6	**Auditability** The availability of metadata that evidences the data quality and information quality factors

Adapted version of Table 1 of Clarke (2016a)

within a strong record of failure, and considerable dispute (e.g. Dreyfus, 1992, Katz, 2012). Successive waves of enthusiasts keep emerging, to frame much the same challenges somewhat differently, and win more grant money from parallel new waves of funding decision-makers. Meanwhile, the water has been muddied by breathless, speculative extensions of AI notions into the realms of metaphysics. In particular, an aside by von Neumann about a 'singularity' has been elevated to spirituality (Moravec, 2000; Kurzweil, 2005), and longstanding sci-fi notions of 'super-intelligence' have been re-presented as philosophy (Bostrom, 2014).

Multiple threads of AI are woven into big data mythology. Various words with a similarly impressive sound to 'intelligent' have been used as marketing banners, such as 'expert', 'neural', 'connectionist', 'learning' and 'predictive'. Definitions are left vague, with each new proposal applying Arthur C. Clarke's Third Law, and striving to be 'indistinguishable from magic' and hence to gain the mantle of 'advanced technology'. Within the research community, expressions of scepticism are in short supply, but Lipton (2015) encapsulates the problem by referring to "an unrealistic expectation that modern feed-forward neural networks exhibit human-like cognition".

One cluster of techniques is marketed as 'machine learning'. A commonly-adopted approach ('supervised learning') involves some kind of (usually quite simple) data structure being provided to a piece of generic software, often one that has an embedded optimisation function. A 'training set' of data is fed in. The process of creating this artefact is claimed to constitute 'learning'. Aspects of the "substantial amount of 'black art'" involved are discussed in Domingos (2012).

Even where some kind of objective is inherent in the data structure and/or the generic software, application of the metaphor of 'learning' is something of stretch for what is a sub-human and in many cases a non-rational process (Burrell, 2016). A thread of work that hopes to overcome some of the weaknesses expands the approach from a single level to a multi-layered model. Inevitably, this too has been given marketing gloss by referring to it as 'deep learning'. Even some enthusiasts are appalled by the hyperbole: "machine learning algorithms [are] not silver bullets, . . . not magic pills, . . . not tools in a toolbox – they are method{ologie}s backed by rational thought processes with assumptions regarding the datasets they are applied to" (Rosebrock, 2014).

A field called 'predictive analytics' over-claims in a different way. Rather than merely extrapolating from a data-series, it involves the extraction of patterns and then extrapolation of the patterns rather than the data; so the claim of 'prediction' is bold. Even some enthusiasts have warned that predictive analytics can have "'unintended side effects' – [things] you didn't really count on when you decided to build models and put them out there in the wild" (Perlich, quoted in Swoyer (2017)).

There is little doubt that there are specific applications to which each particular approach is well-suited – and also little doubt that each is neither a general approach nor deserving of the pretentious title used to market it. As a tweeted aphorism has it: "Most firms that think they want advanced AI/ML really just need linear regression on cleaned-up data" (Hanson, 2016).

The majority of big data analytics activity is performed behind closed doors. One common justification for this is commercial competitiveness, but other factors are commonly at work, in both private and public sector contexts. As a result of the widespread lack of transparency, it is far from clear that practices take into account the many challenges that are identified in this section.

Transparency is in any case much more challenging in the contemporary context than it was in the past. During the early decades of software development, until c.1990, the rationale underlying any particular inference was apparent from the independently-specified algorithm or procedure implemented in the software. Subsequently, so-called expert systems adopted an approach whereby the problem-domain is described, but the problem and solution, and hence the rationale for an inference, are much more difficult to access. Recently, purely empirical techniques such as neural nets and the various approaches to machine learning have attracted a lot of attention. These do not even embody a description of a problem domain. They merely comprise a quantitative summary of some set of instances (Clarke, 1991). In such circumstances, no humanly-understandable rationale for an inference exists, and in many cases none can be created. As a result, transparency is non-existent, and accountability is impossible (Burrell, 2016; Knight, 2017). To cater for such problems, Broeders et al. (2017), writing in the context of national security applications, called for the imposition of a legal duty of care and requirements for external reviews, and the banning of automated decision-making.

This brief review has identified a substantial set of risk factors. Critique is important, but critique is by its nature negative in tone. It is incumbent on critics to also offer positive and sufficiently concrete contributions towards resolution of the problems that they perceive. The primary purpose of this paper is to present a set of Guidelines whose application would address the problems and establish a reliable professional basis for the practice of data analytics.

3. The Guidelines

The Guidelines presented here avoid the word 'big', and refer simply to 'data' and 'data analytics'. These are straightforward and generic terms whose use conveys the prescriptions' broad applicability. The Guidelines are of particular relevance to personal data, because data analytics harbours very substantial threats when applied to data about individuals. The Guidelines are expressed quite generally, however, because inferences drawn from any form of data may have negative implications for individuals, groups, communities, societies, polities, economies or the environment. The purpose of the Guidelines is to assist in the avoidance of harm to all values of all stakeholders. In addition to external stakeholders, shareholders and employees stand to lose where material harm to a company's value arises from poorly-conducted data analytics, including not only financial loss and compliance breaches but also reputational damage.

The Guidelines are presented in Table 2, divided into four segments. Three of the segments correspond to the

Table 2 – Guidelines for the responsible application of data analytics.

1. General

DO's

1.1 **Governance**
Ensure that a comprehensive governance framework is in place prior to, during, and for the relevant period after data acquisition, analysis and use activities, that it is commensurate with the activities' potential impacts, and that it encompasses:
a. risk assessment and risk management from the perspectives of all affected parties
b. express assignments of accountability, at an appropriate level of granularity

1.2 **Expertise**
Ensure that all individuals participating in the activities have education, training, and experience in relation to the real-world systems about which inferences are to be drawn, appropriate to the roles that they play

1.3 **Compliance**
Ensure that all activities are compliant with all relevant laws and established public policy positions within relevant jurisdictions, and with public standards of behaviour

2. Data Acquisition

DO's

2.1 **The Problem Domain**
Understand the real-world systems about which inferences are to be drawn and to which data analytics are to be applied

2.2 **The Data Sources**
Understand each source of data, including:
a. the data's provenance
b. the purposes for which the data was created
c. the meaning of each data-item at the time of creation
d. the data quality at the time of creation
e. the data quality and information quality at the time of use

2.3 **Data Merger**
If data is to be merged from multiple sources, assess the compatibility of the various collections, records and items of data, taking into account the data's provenance, purposes, meaning and quality, and the potential impact of mis-matching and mistaken assumptions

2.4 **Data Scrubbing**
If data is to be scrubbed, cleaned or cleansed, assess the reliability of the processes for the intended purpose and the potential impacts of mistaken assumptions and erroneous changes

2.5 **Identity Protection**
If the association of data with an entity is sensitive, apply techniques to the data whose effectiveness is commensurate with the risks to those entities, in order to ensure pseudonymisation (if the purpose is to draw inferences about individual entities), or de-identification (if the purpose is other than to draw inferences about individual entities)

2.6 **Data Security**
Minimise the risks arising from data acquisition, storage, access, distribution and retention, and manage the unavoidable risks

DON'Ts

2.7 **Identifier Compatibility**
Don't merge data-sets unless the identifiers in each data-set are compatible with one another at a level of reliability commensurate with the potential impact of the inferences drawn

2.8 **Content Compatibility**
Don't merge data-sets unless the reliability of comparisons among the data-items in the sources reaches a threshold commensurate with the potential impact of the inferences drawn

3. Data Analysis

DO's

3.1 **Expertise**
Ensure that all staff and contractors involved in the analysis have:
a. appropriate professional qualifications
b. training in the specific tools and processes
c. sufficient familiarity with the real-world system to which the data relates and with the manner in which the data purports to represent that real-world system
d. accountability for their analyses

3.2 **The Nature of the Tools**
Understand the origins, nature and limitations of data analytic tools that are considered for use

(*continued on next page*)

Table 2 – (*continued*)

3.3 **The Nature of the Data Processed by the Tools**
Understand the assumptions that data analytic tools make about the data that they process, and the extent to which the data to be processed is consistent with those assumptions. Important areas in which assumptions may exist include:
a. the presence of values in relevant data-items
b. the presence of only specific, pre-defined values in relevant data-items
c. the scales against which relevant data-items have been measured
d. the precision with which relevant data-items have been expressed

3.4 **The Suitability of the Tool and the Data**
Demonstrate the applicability of each particular data analytic tool to the particular data that it is proposed be processed using it

DON'Ts

3.5 **Inappropriate Data**
Don't apply data analytics unless the data satisfies threshold tests commensurate with the potential impact of the inferences drawn, in relation to data quality, internal consistency, and reliable correspondence with the real-world systems about which inferences are to be drawn

3.6 **Humanly-Understandable Rationale**
Don't apply an analytical tool that lacks transparency, by which is meant that the rationale for inferences that it draws is expressible in humanly-understandable terms

4. Use of the Inferences

DO's

4.1 **The Impacts**
Understand the potential negative impacts on stakeholders of reliance on the inferences drawn, taking into account the quality of the data and the data analysis process

4.2 **Evaluation**
Where decisions based on inferences from data analytics may have material negative impacts, evaluate the advantages and disadvantages of proceeding, by conducting cost-benefit analysis and risk assessment from an organisational perspective, and impact assessments from the perspectives of other internal and external stakeholders

4.3 **Reality Testing**
Test a sufficient sample of the results of the analysis against the real world, in order to gain insight into the reliability of the data as a representation of relevant real-world entities and their attributes

4.4 **Safeguards**
Design, implement and maintain safeguards and mitigation measures, together with controls that ensure the safeguards and mitigation measures are functioning as intended, commensurate with the potential impacts of the inferences drawn

4.5 **Proportionality**
Where specific decisions based on inferences from data analytics may have material negative impacts on individuals, consider the reasonableness of the decisions prior to committing to them

4.6 **Contestability**
Where actions are taken based on inferences drawn from data analytics, ensure that the rationale for the decisions is transparent to people affected by them, and that mechanisms exist whereby stakeholders can access information about, and if appropriate complain about and dispute interpretations, inferences, decisions and actions

4.7 **Breathing Space**
Provide stakeholders who perceive that they will be negatively impacted by the action with the opportunity to understand and to contest the proposed action

4.8 **Post-Implementation Review**
Ensure that actions and their outcomes are audited, and that adjustments are made to reflect the findings

DON'Ts

4.9 **Humanly-Understandable Rationale**
Don't take actions based on inferences drawn from an analytical tool in any context that may have a material negative impact on any stakeholder unless the rationale for each inference is readily available to those stakeholders in humanly-understandable terms

4.10 **Precipitate Actions**
Don't take actions based on inferences drawn from data analytics until stakeholders who perceive that they may be materially negatively impacted by the action have had a reasonable opportunity to understand and to contest the proposed action. Denial of a reasonable opportunity is only justifiable on the basis of emergency, as distinct from urgency or mere expediency or efficiency. Where a reasonable opportunity is not provided, ensure that stringent safeguards, mitigation measures and controls are designed, implemented and maintained in relation to justification, reporting, review, and recourse in the case of unjustified or disproportionate actions

4.11 **Automated Decision-Making**
Don't delegate to a device any decision that has potentially harmful effects without ensuring that it is subject to specific human approval prior to implementation, by a person who is acting as an agent for the accountable organisation

successive processes involved – acquisition of the data, analysis of the data in order to draw inferences, and use of the inferences. The first segment specifies generic requirements that apply across all of the phases.

Each Guideline is expressed in imperative mode, some in the positive and others in the negative. However, they are not statements of law, nor are they limited to matters that are subject to legal obligations. They are declarations of what is needed in order to manage the risks arising from data quality issues, data meaning uncertainties, incompatibilities in data meaning among similar data-items sourced from different data-collections, misinterpretations of meaning, mistakes introduced by data scrubbing, approaches taken to missing data that may solve some problems but at the cost of creating or exacerbating others, erroneous matches, unjustified assumptions about the scale against which data has been measured, inappropriate applications of analytical tools, lack of review, and confusions among correlation, causality, predictive power and normative force.

The organisations and individuals to whom each Guideline is addressed will vary depending on the context. In some circumstances, a single organisation, a single small team within an organisation, or even a single individual, might perform all of the activities involved. On the other hand, multiple teams within one organisation, or across multiple organisations, may perform several of the activities.

The Guidelines are intended to be comprehensive. As a result, in any particular context, some of them will be redundant, and some would be more usefully expressed somewhat differently. In particular, some of the statements are primarily relevant to data that refers to an individual human being. Such statements may be irrelevant, or may benefit from re-phrasing, where the data relates to inanimate parts of the physical world (e.g. meteorological, geophysical, vehicular traffic or electronic traffic data), or to aggregate economic or social phenomena. In such circumstances, careful sub-setting and adaptation of the Guidelines is appropriate.

4. Ways to apply the Guidelines

These Guidelines, in their current or some adapted form, can be adopted by any organisation. Staff and contractors can be required to demonstrate that their projects are compliant, or, to the extent that they are not, to explain why not. In practice, adoption may be driven by staff and contractors, because many practitioners are concerned about the implications of their work, and would welcome the availability of an instrument that enables them to raise issues in the context of project risk management.

Organisational self-regulation of this kind has the capacity to deliver value for the organisation and for shareholders, but it has only a mediocre track-record in benefiting stakeholders outside the organisation. A stronger institutional framework is needed if preventable harm arising from inappropriate data, analysis and use is to be avoided.

Industry associations can adopt or adapt the Guidelines, as can government agencies that perform oversight functions. Industry regulation through a Code of Practice may achieve some positive outcomes for organisations in terms of the quality of work performed, and particularly by providing a means of defending against and deflecting negative media reports, public concerns about organisational actions, and acts by any regulator that may have relevant powers. In practice, however, such Codes are applied by only a proportion of the relevant organisations, are seldom taken very seriously (such as by embedding them within corporate policies, procedures, training programs and practices), are unenforceable, and generally offer very limited benefits to external stakeholders. Nonetheless, some modest improvements would be likely to accrue from adoption, perhaps at the level of symbolism, but more likely as a means of making it more difficult for data analytics issues to be ignored.

Individual organisations can take positive steps beyond such, largely nominal, industry sector arrangements. They can embed consideration of the factors identified in these Guidelines into their existing business case, cost/benefit and/or risk assessment and management processes. In order to fulfil their corporate social responsibility commitments, they can also evaluate proposed uses of data analytics from the perspectives of external stakeholders. A very narrow and inadequate approach to this merely checks legal compliance, as occurs with the pseudo-PIA processes conventional in the public sector throughout much of North America (Clarke, 2011 s.4), and in the new European 'Data Protection Impact Assessment' (DPIA) mechanism (Clarke, 2017a). Much more appropriately, a comprehensive Privacy Impact Assessment can be performed (Clarke, 2009; Wright and de Hert, 2012). In some circumstances, a much broader social impact assessment is warranted (Raab and Wright, 2012; Wright and Friedewald, 2013). Raab & Wright (2012 pp. 379–381) calls for extension of the scope of PIAs firstly to a wide range of impacts on the individual's "relationships, positions and freedoms", then to "impacts on groups and categories", and finally to "impacts on society and the political system".

A further step that individual organisations can take is to enter into formal undertakings to comply with a Code, combined with submission to the decisions of a complaints body, ombudsman or tribunal that is accessible by any aggrieved party that has the resources to conduct investigations, that has enforcement powers, and that uses them. Unfortunately, such arrangements are uncommon, and it is not obvious that suitable frameworks exist within which an enforceable Code along the lines of these Guidelines could be implemented.

Another possibility is for a formal and sufficiently precise Standard to be established, and for this to be accepted by courts as the measuring-stick against which the behaviour of organisations that conduct data analytics is to be measured. A loose mechanism of this kind is declaration by an organisation that it is compliant with a particular published Standard. In principle, this would appear to create a basis for court action by aggrieved parties. In practice, however, it appears that such mechanisms are seldom effective in protecting either internal or external stakeholders.

As discussed earlier, some documents exist that at least purport to provide independent guidance in relation to data analytics activities. These Guidelines can be used as a yardstick against which such documents can be measured. The UK Cabinet Office's 'Data Science Ethical Framework' (UKCO,

2016) was assessed against an at-that-time-unformalised version of these Guidelines, and found to be seriously wanting (Raab and Clarke, 2016). For different reasons, and in different ways, the Council of Europe document (CoE 2017) falls a very long way short of what is needed by professionals and the public alike as a basis for responsible use of data analytics. The US Government Accountability Office has identified the existence of "possible validity problems in the data and models used in [data analytics and innovation efforts – DAI]" (GAO, 2016, p. 38), but has done nothing about them. An indication of the document's dismissiveness of the issues is this quotation: "In automated decision making [using machine learning], monitoring and assessment of data quality and outcomes are needed to gain and maintain trust in DAI processes" (p.13, fn.8). Not only does the statement appear in a mere footnote, but the concern is solely about 'trust' and not at all about the appropriateness of the inferences drawn, the actions taken as a result of them, or the resource efficiency and equitability of those actions. The current set of documents from the US National Institute of Standards and Technology (NIST, 2015) is also remarkably devoid of discussion about data quality and process quality, and offers no process guidance along the lines of the Guidelines proposed in this paper.

Another avenue whereby progress can be achieved is through adoption by the authors of text-books. At present, leading texts commonly have a brief, excusatory segment, usually in the first or last chapter. Curriculum proposals commonly suffer the same defect, e.g. Gupta et al. (2015), Schoenherr and Speier-Pero (2015). Course-designers appear to generally follow the same pattern, and schedule a discussion or a question in an assignment, which represents a sop to the consciences of all concerned, but does almost nothing about addressing the problems, and nothing about embedding solutions to those problems within the analytics process. It is essential that the specifics of the Guidelines in Table 2 be embedded in the structure of text-books and courses, and that students learn to consider each issue at the point in the acquisition/analysis/use cycle at which each challenge needs to be addressed.

None of these approaches is a satisfactory substitute for legislation that places formal obligations on organisations that apply data analytics, and that provides aggrieved parties with the capacity to sue organisations where they materially breach requirements and there are material negative impacts. Such a scheme may be imposed by an activist legislature, or a regulatory framework may be legislated and the Code negotiated with the relevant parties prior to promulgation by a delegated agency. It is feasible for organisations themselves to submit to a parliament that a co-regulatory scheme of such a kind should be enacted, for example where scandals arise from inappropriate use of data analytics by some organisations, which have a significant negative impact on the reputation of an industry sector as a whole.

5. Conclusions

This paper has not argued that big data and big data analytics are inherently evil. It has also not argued that no valid applications of the ideas exist, nor that all data collections are of such low quality that no useful inferences can be drawn from them, nor that all mergers of data from multiple sources are necessarily logically invalid or necessarily deliver fatally flawed consolidated data-sets, nor that all data scrubbing fails to clean data, nor that all data analytics techniques make assumptions about data that can under no circumstances be justified, nor that all inferences drawn must be wrong. Expressed in the positive, some big data has potential value, and some applications of data analytics techniques are capable of realising that potential.

What this paper has done is to identify a very large fleet of challenges that have to be addressed by each and every specific proposal for the expropriation of data, the re-purposing of data, the merger of data, the scrubbing of data, the application of data analytics to it, and the use of inferences drawn from the process in order to make, or even guide, let alone explain, decisions and action that affect the real world. Further, it is far from clear that measures are being adopted to meet these challenges.

Ill-advised applications of data analytics are preventable by applying the Guidelines proposed in this paper. As the 'big data' mantra continues to cause organisations to have inflated expectations of what data analytics can deliver, both shareholders and external stakeholders need constructive action to be taken in order to get data analytics practices under control, and avoid erroneous business decisions, loss of shareholder value, inappropriate policy outcomes, and unjustified harm to individual, social, economic and environmental values. The Guidelines proposed in this paper therefore provide a basis for the design of organisational and regulatory processes whereby positive benefits can be gained from data analytics, but undue harm avoided.

Acknowledgement

The author received valuable feedback from Prof. Louis de Koker of La Trobe University, Melbourne, David Vaile and Dr. Lyria Bennett Moses of UNSW, Sydney, Dr. Kerry Taylor of the ANU, Canberra, Dr. Kasia Bail of the University of Canberra, Prof. Charles Raab of Edinburgh University, and an anonymous reviewer. Evaluative comments are those of the author alone.

REFERENCES

ACM. Statement on algorithmic transparency and accountability. Association for Computing Machinery; 2017. Available from: https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf. [Accessed November 24, 2017].

Agrawal D, Philip Bernstein P, Bertino E, Davidson S, Dayal U, Franklin M, Gehrke J, et al. Challenges and opportunities with big data 2011-1. Cyber Center Technical Reports, Paper 1; 2011. Available from: http://docs.lib.purdue.edu/cctech/1. [Accessed November 24, 2017].

Anderson C. The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine 16:07; 2008.

ASA. Ethical guidelines for statistical practice. American Statistical Association; 2016. Available from: http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx. [Accessed November 24, 2017].
Bollier D. The promise and peril of big data. The Aspen Institute; 2010. Available from: https://www.emc.co.tt/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf. [Accessed November 24, 2017].
Bostrom N. Superintelligence: paths, dangers, strategies. Oxford University Press; 2014.
Boyd D, Crawford K. Six provocations for big data. Proc. Symposium on the Dynamics of the Internet and Society; 2011. Available from: http://ssrn.com/abstract=1926431. [Accessed November 24, 2017].
Broeders D, Schrijvers E, van der Sloot B, van Brakel R, de Hoog J, Ballina EH. Big data and security policies: towards a framework for regulating the phases of analytics and use of big data. Comput Law Secur Rev 2017;33:309–23.
Burrell J. How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc 2016;3(1):1–12.
Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data Sci J 2015;14(2):1–10. Available from: https://datascience.codata.org/articles/10.5334/dsj-2015-002/.
Cao L. Data science: a comprehensive overview. ACM Computing Surveys; 2017. Available from: http://dl.acm.org/ft_gateway.cfm?id=3076253&type=pdf. [Accessed November 24, 2017].
Clarke R. A contingency approach to the software generations. Database 1991;22(3):23–34. PrePrint available from: http://www.rogerclarke.com/SOS/SwareGenns.htmlSummer 1991.
Clarke R. Dataveillance by governments: the technique of computer matching. Inf Tech People 1994;7(2):46–85. PrePrint available from: http://www.rogerclarke.com/DV/MatchIntro.html.
Clarke R. Privacy impact assessment: its origins and development. Comput Law Secur Rev 2009;25(2):123–35. PrePrint available from: http://www.rogerclarke.com/DV/PIAHist-08.html.
Clarke R. An evaluation of privacy impact assessment guidance documents. Int Data Priv Law 2011;1(2):111–20. PrePrint available from: http://www.rogerclarke.com/DV/PIAG-Eval.html.
Clarke R. Big data, big risks. Inf Syst J 2016a;26(1):77–90. PrePrint available from: http://www.rogerclarke.com/EC/BDBR.html.
Clarke R. Quality assurance for security applications of big data. Proc. European Intelligence and Security Informatics Conference (EISIC), Uppsala, 17–19 August 2016; 2016b. PrePrint available from: http://www.rogerclarke.com/EC/BDQAS.html. [Accessed November 24, 2017].
Clarke R. The distinction between a PIA and a Data Protection Impact Assessment (DPIA) under the EU GDPR. Working Paper, Xamax Consultancy Pty Ltd; 2017a. Available from: http://www.rogerclarke.com/DV/PIAvsDPIA.html. [Accessed November 24, 2017].
Clarke R. Big data prophylactics, chapter 1. In: Lehmann A, Whitehouse D, Fischer-Hübner S, Fritsch L, Raab C, editors. Privacy and identity management. Facing up to next steps. Springer; 2017b. p. 3–14 [chapter 1]. PrePrint available from: http://www.rogerclarke.com/DV/BDP.html.
CoE. Guidelines on the protection of individuals with regard to the processing of personal data in a world of big data. Convention 108 Committee, Council of Europe; 2017. Available from: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806ebe7a. [Accessed November 24, 2017].
DoFD. Better practice guide for big data. Australian Dept of Finance & Deregulation, v.2; 2015. Available from: http://www.finance.gov.au/sites/default/files/APS-Better-Practice-Guide-for-Big-Data.pdf. [Accessed November 24, 2017].
Domingos P. A few useful things to know about machine learning. Commun ACM 2012;55(10):78–87.
Dreyfus H. What computers still can't do. MIT Press; 1992.
DSA. Data science code of professional conduct. Data Science Association, undated but apparently of 2016; 2016. Available from: http://www.datascienceassn.org/sites/default/files/datasciencecodeofprofessionalconduct.pdf. [Accessed November 24, 2017].
GAO. Emerging opportunities and challenges data and analytics innovation. Government Accountability Office, Washington DC; 2016. Available from: http://www.gao.gov/assets/680/679903.pdf. [Accessed November 24, 2017].
Gupta B, Goul M, Dinter B. Business intelligence and big data in higher education: status of a multi-year model curriculum development effort for business school undergraduates, MS graduates, and MBAs. Commun Assoc Inf Syst 2015;36(23):Available from: https://www.researchgate.net/profile/Babita_Gupta4/publication/274709810_Communications_of_the_Association_for_Information_Systems/links/557ecd4b08aeea18b7795225.pdf.
Hanson R. This AI boom will also bust. Overcoming Bias Blog; 2016. Available from: http://www.overcomingbias.com/2016/12/this-ai-boom-will-also-bust.html. [Accessed November 24, 2017].
Haryadi AF, Hulstijn J, Wahyudi A, van der Voort H, Janssen M. Antecedents of big data quality: an empirical examination in financial service organizations. Proc. IEEE Int'l Conf. on Big Data; 2016. pp. 116–21. Available from: https://pure.tudelft.nl/portal/files/13607440/Antecedents_of_Big_Data_Quality_IEEE2017_author_version.pdf. [Accessed November 24, 2017].
Hazen BT, Boone CA, Ezell JD, Jones-Farmer LA. Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications. Int J Prod Econ 2014;154:72–80. Available from: https://www.researchgate.net/profile/Benjamin_Hazen/publication/261562559_Data_Quality_for_Data_Science_Predictive_Analytics_and_Big_Data_in_Supply_Chain_Management_An_Introduction_to_the_Problem_and_Suggestions_for_Research_and_Applications/links/0deec534b4af9ed874000000.
Huh YU, Keller FR, Redman TC, Watkins AR. Data quality. Inf Softw Tech 1990;32(8):559–65.
ICO. Big data, artificial intelligence, machine learning and data protection. UK Information Commissioner's Office, Discussion Paper v.2.2; 2017. Available from: https://ico.org.uk/for-organisations/guide-to-data-protection/big-data/. [Accessed November 24, 2017].
Jagadish HV. Big data and science: myths and reality. Big Data Res 2015;2(2):49–52.
Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, et al. Big data and its technical challenges. Commun ACM 2014;57(7):86–94.
Kandel S, Heer J, Plaisant C, Kennedy J, van Ham F, Henry-Riche N, et al. Research directions for data wrangling: visualizations and transformations for usable and credible data. Information Visualization 10. 4; 2011. 271–88. Available from: https://idl.cs.washington.edu/files/2011-DataWrangling-IVJ.pdf. [Accessed November 24, 2017].
Katz Y. Noam Chomsky on where artificial intelligence went wrong: an extended conversation with the legendary linguist. The Atlantic; 2012. Available from: https://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/. [Accessed November 24, 2017].

King NJ, Forder J. Data analytics and consumer profiling: finding appropriate privacy principles for discovered data. Comput Law Secur Rev 2016;32:696–714.
Knight W. The dark secret at the heart of AI. 11 April 2017, MIT Technology Review; 2017. Available from: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/. [Accessed November 24, 2017].
Kurzweil R. The singularity is near: when humans transcend biology. Viking; 2005.
Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in big data analysis. Science 2014;343(6176):1203–5. Available from: https://dash.harvard.edu/bitstream/handle/1/12016836/The%20Parable%20of%20Google%20Flu%20%28WP-Final%29.pdf.
Lipton ZC. (Deep Learning's Deep Flaws)'s Deep Flaws. KD Nuggets; 2015. Available from: http://www.kdnuggets.com/2015/01/deep-learning-flaws-universal-machine-learning.html. [Accessed November 24, 2017].
Mayer-Schonberger V, Cukier K. Big data, a revolution that will transform how we live, work and think. John Murray; 2013.
McFarland DA, McFarland HR. Big data and the danger of being precisely inaccurate. Big Data Soc 2015;2(2):1–4.
Merino J, Caballero I, Bibiano R, Serrano M, Piattini M. A data quality in use model for big data. Fut Gen Comput Syst 2016;63:123–30.
Metcalf J, Crawford K. Where are human subjects in big data research? The emerging ethics divide. Big Data Soc 2016;3(1):1–14.
Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. Big Data Soc 2016;3(2):1–21.
Moravec H. Robot: mere machine to transcendent mind. Oxford University Press; 2000.
Müller H, Freytag J-C. Problems, methods and challenges in comprehensive data cleansing. Technical Report HUB-IB-164, Humboldt-Universität zu Berlin, Institut für Informatik; 2003. Available from: http://www.informatik.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/MuFr03.pdf. [Accessed November 24, 2017].
Müller O, Junglas I, vom Brocke1 J, Debortoli S. Utilizing big data analytics for information systems research: challenges, promises and guidelines. Eur J Inf Syst 2016;25(4):289–302. Available from: https://www.researchgate.net/profile/Oliver_Mueller5/publication/290973859_Utilizing_Big_Data_Analytics_for_Information_Systems_Research_Challenges_Promises_and_Guidelines/links/56ec168f08aee4707a384fff/Utilizing-Big-Data-Analytics-for-Information-Systems-Research-Challenges-Promises-and-Guidelines.pdf.
NIST. NIST big data interoperability framework. Special Publication 1500-1, v.1, National Institute of Standards and Technology; 2015. Available from: https://bigdatawg.nist.gov/V1_output_docs.php. [Accessed November 24, 2017].
OAIC. Consultation draft: guide to big data and the Australian Privacy Principles. Office of the Australian Information Commissioner; 2016. Available from: https://www.oaic.gov.au/engage-with-us/consultations/guide-to-big-data-and-the-australian-privacy-principles/consultation-draft-guide-to-big-data-and-the-australian-privacy-principles. [Accessed November 24, 2017].
Piprani B, Ernst D. A model for data quality assessment. Proc. OTM Workshops (5333); 2008. pp 750–9.
Press G. A very short history of data science. Forbes; 2013. Available from: https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#375c75e355cf. [Accessed November 24, 2017].
Raab C, Clarke R. Inadequacies in the UK's data science ethical framework. Euro Data Protect L 2016;2(4):555–60. PrePrint available from: http://www.rogerclarke.com/DV/DSEFR.html.
Raab CD, Wright D, de Hert P. editors. Surveillance: extending the limits of privacy impact assessment. 2012. p. 363–83 [Ch. 17].
Rahm E, Do HH. Data cleaning: problems and current approaches. IEEE Data Eng Bull 2000;23:Available from: http://dc-pubs.dbs.uni-leipzig.de/files/Rahm2000DataCleaningProblemsand.pdf.
Rivers CM, Lewis BL. Ethical research standards in a world of big data. F1000Res 2014;3:38. Available from: https://f1000research.com/articles/3-38.
Rosebrock A. Get off the deep learning bandwagon and get some perspective. PY Image Search; 2014. Available from: https://www.pyimagesearch.com/2014/06/09/get-deep-learning-bandwagon-get-perspective/. [Accessed November 24, 2017].
Saha B, Srivastava D. Data quality: The other face of big data. Proc. Data Engineering (ICDE); 2014. pp. 1294–7. Available from: https://people.cs.umass.edu/~barna/paper/ICDE-Tutorial-DQ.pdf. [Accessed November 24, 2017].
Schoenherr T, Speier-Pero C. Data science, predictive analytics, and big data in supply chain management: current state and future potential. J Bus Logist 2015;36(1):120–32. Available from: http://www.logisticsexpert.org/top_articles/2016/2016%20-%20Research%20-%20JBL%20-%20Data%20Science,%20Predictive%20Analytics,%20and%20Big%20Data%20in%20Supply%20Chain%20Managementl.pdf.
Shanks G, Darke P. Understanding data quality in a data warehouse. Aust Comput J 1998;30:122–8.
Swoyer S. The shortcomings of predictive analytics. TDWI; 2017. Available from: https://tdwi.org/articles/2017/03/08/shortcomings-of-predictive-analytics.aspx. [Accessed November 24, 2017].
UKCO. Data science ethical framework. U.K. Cabinet Office, v.1.0; 2016. Available from: https://www.gov.uk/government/publications/data-science-ethical-framework. [Accessed November 24, 2017].
UNSD. Declaration of professional ethics. United Nations Statistical Division; 1985. Available from: http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=93#start. [Accessed November 24, 2017].
Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 1996;12(4):5–33. Spring, 1996.
Wigan MR, Clarke R. Big data's big unintended consequences. IEEE Comput 2013;46(6):46–53. PrePrint available from: http://www.rogerclarke.com/DV/BigData-1303.html.
WP29. Statement of the WP29 on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU. Article 29 Working Party, European Union; 2014. Available from: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp221_en.pdf. [Accessed November 24, 2017].
Wright D, de Hert P, editors. Privacy impact assessments. Springer; 2012.
Wright D, Friedewald M. Integrating privacy and ethical impact assessments. Sci Public Policy 2013;40(6):755–66. Available from: http://spp.oxfordjournals.org/content/40/6/755.full.
Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, Goodman A, et al. Ten simple rules for responsible big data research. PLoS Comput Biol 2017;13(3):Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5373508/. [Accessed November 24, 2017].

Guidelines for the Responsible Business Use of AI
Foundational Working Paper

Stable Version of 3 October 2018
(Added ss 5.1 and 5.9, additional citations, editorials, re-formatting)
(Substantial further development of s.7, addition of Table 4 and Appendices)

Roger Clarke **

This document is at http://www.rogerclarke.com/EC/GAIF.html

Abstract

Organisations across the private and public sectors are looking to use artificial intelligence (AI) techniques not only to draw inferences, but also to make decisions and take action, and even to do so autonomously. This is despite the absence of any means of programming values into technologies and artefacts, and the obscurity of the rationale underlying inferencing using contemporary forms of AI.

To what extent is AI really suitable for real-world applications? Can corporate executives satisfy their board-members that the business is being managed appropriately if AI is inscrutable? Beyond operational management, there are compliance risks to manage, and threats to important relationships with customers, staff, suppliers and the public. Ill-advised uses of AI need to be identified in advance and nipped in the bud, to avoid harm to important values, both corporate and social. Organisations need to extract the achievable benefits from advanced technologies rather than dreaming dangerous dreams.

This working paper first considers several approaches to addressing the gap between the current round of AI marketing hype and the hard-headed worlds of business and government. It is first proposed that AI needs to be re-conceived as 'complementary intelligence', and that the robotics notion of 'machines that think' needs to give way to the idea of 'intellectics', with the focus on 'computers that do'.

A review of 'ethical analysis' of IT's impacts extracts little of value. A consideration of regulatory processes proves to be of more use, but to still deliver remarkably little concrete guidance. It is concluded that the most effective approach for organisations to take is to apply adapted forms of the established techniques of risk assessment and risk management. Critically, stakeholder analysis needs to be performed, and risk assessment undertaken, from those perspectives as well as from that of the organisation itself. This Working Paper's final contribution is to complement that customised form of established approaches to risk by the presentation of a derivative set of Principles for Responsible AI, with indications provided of how those Principles can be operationalised for particular forms of complementary intelligence and intellectics.

Contents

1. Introduction

The term Artificial Intelligence (AI) was coined in 1955 in a proposal for the 1956 Dartmouth Summer Research Project in Automata (McCarthy et al. 1955). The proposal was based on "the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it". Histories of AI (e.g. Russell & Norvig 2009, pp. 16-28) identify multiple strands, but also multiple re-visits to much the same territory, and a considerable

degree of creative chaos.

The over-enthusiasm that characterises the promotion of AI has deep roots. Simon (1960) averred that "Within the very near future - much less than twenty-five years - we shall have the technical capability of substituting machines for any and all human functions in organisations. ... Duplicating the problem-solving and information-handling capabilities of the brain is not far off; it would be surprising if it were not accomplished within the next decade". Over 35 years later, with his predictions abundantly demonstrated as being fanciful, Simon nonetheless maintained his position, e.g. "the hypothesis is that a physical symbol system [of a particular kind] has the necessary and sufficient means for general intelligent action" (Simon 1996, p. 23 - but expressed in similar terms from the late 1950s, in 1969, and through the 1970s), and "Human beings, viewed as behaving systems, are quite simple" (p. 53). Simon acknowledged "the ambiguity and conflict of goals in societal planning" (p. 140), but his subsequent analysis of complexity (pp. 169-216) considered only a very limited sub-set of the relevant dimensions. Much the same dubious assertions can be found in, for example, Kurzweil (2005): "by the end of the 2020s" computers will have "intelligence indistinguishable to biological humans" (p.25), and in self-promotional documents of the current decade.

AI has offered a long litany of promises, many of which have been repeated multiple times, on a cyclical basis. Each time, proponents have spoken and written excitedly about prospective technologies, using descriptions that not merely verged into the mystical, but often crossed the border into the realms of magic and alchemy. Given the habituated exaggerations that proponents indulge in, it is unsurprising that the field has exhibited cyclical 'boom and bust' patterns, with research funding being sometimes very easy to obtain, and sometimes very difficult, depending on whether the focus at the time is on the hyperbole or on the very low delivery-rate against promises.

Part of AI's image-problem is that most of the successes deriving from what began as AI research have shed the name, and become associated with other terms. For example, pattern recognition, variously within text, speech and two-dimensional imagery, has made a great deal of progress, and achieved application in multiple fields, as diverse as dictation, vehicle number-plate recognition and object and facial recognition. Expert systems approaches, particularly based on rule-sets, have also achieved a degree of success. Game-playing, particularly of chess and go, have surpassed human-expert levels, and provided entertainment value and spin-offs, but seem not to have provided the breakthroughs towards posthumanism that their proponents appeared to be claiming for them.

This Working Paper concerns itself with the question of how organisations can identify AI technologies that have practical value, and apply them in ways that achieve benefits, without incurring disproportionate disbenefits or giving rise to unjustified risks. A key feature of AI successes to date appears to be that, even where the technology or its application is complex, it is understandable by people with appropriate technical background, i.e. it is not magic and is not presented as magic, and its applications are auditable. AI technologies that have been effective have been able to be empirically tested in real-world contexts, but under sufficiently controlled conditions that the risks have been able to be managed.

The scope addressed in this Working Paper is very broad, in terms of both technologies and applications, but it does not encompass design and use for warfare or armed conflict. It does, however, include applications to civil law enforcement and domestic national security, i.e. safeguards for the public, for infrastructure, and for public figures.

This working paper commences by considering interpretations of the AI field that may contribute to overcoming its problems and assist in analysing the opportunities and threats that it embodies. Brief scans are undertaken of current technologies that are within the field of view. There are several possible sources of guidance in relation to the responsible use of AI. The paper first considers ethics, and then regulatory regimes. It proposes, however, that the most useful approach is through risk assessment and management processes, but expanding the perspectives from solely that of the organisation itself to also embrace those of stakeholders. The final section draws on the available sources in order to propose a set of principles for the responsible application of AI that are specific enough to guide organisations' business processes.

---

2. Rethinking AI

A major contributor to AI's problems has been the diverse and often conflicting conceptions of what it is, and what it is trying to achieve. The first necessary step is to disentangle the key ideas, and adopt an interpretation that can assist user organisations to appreciate the nature of the technology, and then analyse its potential contributions and downsides.

2.1 'AI' cf. 'Complementary Intelligence'

What does, what could, and what should 'intelligence' mean? What does 'artificial' mean? And are the conventional interpretations of these terms useful to individual organisations, and to the economy and society more generally?

The general sense in which the term 'intelligence' is used by the AI community is that an entity exhibits intelligence if it has 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). Some AI proponents strive to replicate in artefacts the processes whereby human entities exhibit intelligence, whereas others define AI in terms of the artefact's performance rather than the means whereby the performance arises.

The term 'artificial' has always been problematic. The originators of the term used it to mean 'synthetic', in the sense of being human-made but equivalent to human. It is far from clear that there was a need for yet more human intelligence in 1955, when there were 2.8 billion people, let alone now, when there are over 7 billion of us, many under-employed and likely to remain so.

Some proponents have shifted away from human-equivalence, and posited that AI is synthetic, but in some way 'superior-to-human'. This raises the question as to how superiority is to be measured. For example, is playing a game better than human experts necessarily a useful measure? There is also a conundrum embedded in this approach: if human intelligence is inferior, how can it reliably define what 'superior-to-human' means?

An alternative approach may better describe what humankind needs. An idea that is traceable at least to Wyndham (1932) is that " ... man and machine are natural complements: They assist one another". I argued in Clarke (1989) that there was a need to "deflect the focus ... toward the concepts of 'complementary intelligence' and 'silicon workmates' ... to complement human strengths and weaknesses, rather than to compete with them". Again, in Clarke (1993), reprised in Clarke (2014b), I reasoned that: "Because robot and human capabilities differ, for the foreseeable future at least, each will have specific comparative advantages. Information technologists must delineate the relationship between robots and people by applying the concept of decision structuredness to blend computer-based and human elements advantageously".

Adopting this approach, AI needs to be re-conceived such that its purpose is to extend human capabilities. Rather than 'artificial' intelligence, the design objective needs to be 'complementary' intelligence, the essence of which is:

1. to do things well that humans do badly or cannot do at all; and
2. to function as elements within systems that include both humans and artefacts, with effective, efficient and adaptable interfacing among them all.

An important category of 'complementary intelligence' is the use of negative-feedback mechanisms to achieve automated equilibration within human-made systems. A longstanding example is the maintenance of ship trim and stability by means of hull shape and careful weight distribution, including ballast. A more commonly celebrated instance is Watts' fly-ball governor for regulating the pressure in a boiler. Of more recent origin are schemes to achieve real-time control over the orientation of craft floating in fluids, and maintenance of their location or path. There are successful applications to deep-water oil-rigs, underwater craft, and aircraft both with and without pilots on board. The notion is also exemplified by the distinction between decision support systems (DSS), which are designed to assist humans make decisions, and decision systems (DS), whose purpose is to make the decisions without human involvement.

Computer-based systems have a clear advantage over humans in contexts in which significant computation is involved, reliability and accuracy are important, and speed of inferencing, decision-making and/or action-taking, are important. This advantage is, however, limited to circumstances in which either a structured process exists or heuristics or purely empirical techniques have been well-demonstrated to be effective.

Further advantages may arise in relation to cost, the delegation to devices of boringly mundane tasks, and the performance by artefacts of tasks that are inherently dangerous, or that need to be performed in environments that are inherently dangerous to humans and/or are beyond their physical capabilities (e.g. environments that feature high pressure such as deep water, low pressure such as space, or high radiation levels both in space and close to nuclear materials). Even where such superiority can be demonstrated, however, the need exists to focus discussion about AI on complementary intelligence, on technologies that augment human capabilities, and on systems that feature collaboration between humans and artefacts.

I contend that the use of the complementary intelligence notion can assist organisations in their efforts to distinguish uses of AI that have prospects for adoption, the generation of net benefits, the management of disbenefits, and the achievement of public acceptability.

2.2 Autonomy

The concept of 'automation' is concerned with the performance of a predetermined procedure, or response in predetermined ways to alternative stimuli. It is observable in humans, e.g. under hypnosis, and is designed-into many kinds of artefacts.

The rather different notion of 'autonomy' means, in humans, the capacity for independent decision and action. Further, in some contexts, it also encompasses a claim to the right to exercise that capacity. It is associated with the notions of consciousness, sentience, self-awareness, free will and self-determination. Autonomy in artefacts, on the other hand, lies much closer to the notion of automation. It may merely refer to a substantial repertoire of pre-programmed stimulus-response relationships. Alternatively, it may refer to some degree of adaptability to context, as might arise if some form of machine-learning were included, such that the specification of the stimulus-response relationships change over time depending on the cases handled in the intervening period. Another approach might be to define artefact autonomy in terms of the extent to which a human or some other artefact, does, or even can, intervene in the artefact's behaviour.

In humans, autonomy is best approached as a layered phenomenon. Each of us performs many actions in a subliminal manner. For example, our eye and ear receptors function without us ever being particularly aware of them, and several layers of our neural systems handle the signals in order to offer us cognition, that is to say awareness and understanding, of the world around us.

A layered approach is applicable to artefacts as well. Aircraft generally, including drones, may have layers of behaviour that occur autonomously, without pilot action or even awareness. Maintenance of the aircraft's 'attitude' (orientation to the vertical and horizontal), and angle to the wind-direction, may, from the pilot's viewpoint, simply happen. At a higher level of delegation, the aircraft may adjust the aircraft's flight controls in order to maintain a predetermined flight-path, and in the case of rotorcraft, to maintain the vehicle's location relative to the earth's surface. A higher-order autonomous function is inflight manoeuvring to avoid collisions. At a yet higher level, some aircraft can perform take-off and/or landing autonomously. To date, high-order activities that are seldom if ever autonomous include decisions about when to take off and land, the mission objective, and 'target acquisition' (where to land, where to deliver a payload, which location to direct the payload towards).

At the lower levels, the rapidity with which analysis, decision and action need to be taken may preclude conscious human involvement. At the higher levels, however, a pilot may be able to request advice, to accept or reject advice, to authorise an action recommended by an artefact, to override or countermand a default action, or to resume full control. From the perspective of the drone, its functions may be to perform until its autonomous function is revoked, to perform except where a particular action is over-ridden, to recommend, to advise, or to do nothing.

IEEE, even thought it is one of the most relevant professional associations in the field, made no meaningful attempt to address these issues for decades. It is currently endeavouring to do so. It commenced with a discussion paper (IEEE 2017) which avoids the term AI, and instead uses the term 'Autonomous and Intelligent Systems (A/IS)'. This highlights the need to address both intelligence and autonomy in an integrated manner.

2.3 Technology, Artefacts, Systems and Applications

A further factor that has tended to cloud meaningful discussion of responsibility in relation to AI has been 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 various categories of entities that are active in each phase.

IEEE's discussion paper (IEEE 2017) recognises that the end-result of successive rounds of R&D is complex systems that are applied in real-world contexts. In order to deliver such systems, however, technology has to be conceived, proven, and embedded in artefacts. It is therefore valuable to distinguish between technology, artefacts that embody the technology, systems that incorporate the artefacts, and applications of those systems. Appropriate responsibilities can then be assigned to researchers, to inventors, to innovators, to purveyors, and to users. Table 1 identifies phases, the output from each phase, and the categories of entity that bear legal and moral responsibility for disbenefits arising from AI.

Table 1: Entities with Responsibilities in Relation to AI

Phase	Result	Responsibility
Research	AI Technology	Researchers
Invention	AI-Based Artefacts	R&D Engineers
Innovation	AI-Based Systems	Developers
Dissemination	Installed AI-Based Systems	Purveyors
Application	Impacts	User Organisations and Individuals

This section has proposed several measures whereby the fog induced by the AI notion can be lifted, and a framework developed for managing AI-based activities. The focus needs to be on complementary intelligence and autonomy, as features of technology, artefacts, systems and applications that support collaboration among all system elements.

---

3. Contemporary AI

AI's scope is broad, and contested. This section identifies areas that have current relevance. Their relevance derives in part from claims of achievement of progress and benefits, and in part from media coverage resulting in awareness among both organisations' staff and the general public. In addition to achieving some level of adoptiong, each faces to at least some degree technical challenges, public scepticism and resistance. Achievement of the benefits that are potentially extractable from these technologies is also threatened by over-claiming, over-reach, and resulting loss of public confidence. This section considers three forms of AI, and then suggests an alternative conceptualisation intended to assist in understanding and addressing the technical, acceptance and adoption challenges.

3.1 Robotics

Robotics originally emerged in the form of machines enhanced with computational capacity. The necessary elements are sensors to acquire data from the robot's environment, computing hardware and software to enable inferences to be drawn and decisions made, and actuators in order to give effect to those decisions by acting on the robot's environment. Robotics has enjoyed its major areas of success in controlled environments such as the factory floor and the warehouse. Less obviously 'robotic' systems include low-level control over the attitude, position and course of craft on or in water and in the air.

The last few years have seen a great deal of coverage of self-driving vehicles, variously on rails and otherwise, in controlled environments such as mines and quarries and dedicated bus routes, and recently in more open environments. In addition, robotics has taken flight, in the form of drones (Clarke 2014a).

Many claims have been made recently about 'the Internet of Things' (IoT) and about systems comprising many small artefacts, such as 'smart houses' and 'smart cities'. For a consolidation and rationalisation of multiple such ideas into the notion of an 'eObject', see Manwaring & Clarke (2015). Many of the initiatives in this area are robotic in nature, in that they encompass all of sensors, computing and actuators.

3.2 Cyborgisation

The term cyborgisation refers to the process of enhancing individual humans by technological means, such that a cyborg is a hybrid 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 that are used to enhance humans include 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.

Many such artefacts replace lost functionality, and are referred to as prosthetics. Others, which can be usefully referred to as orthotics, provide augmented or additional functionality (Clarke 2011). An example of an orthotic is augmented reality for firefighters, displaying building plans and providing object-recognition in their visual field. It was argued in Clarke (2014b) that use by drone pilots of instrument-based remote control, and particularly of first-person view (FPV) headsets, represent a form of orthotic cyborgisation.

Artefacts of these kinds are not commonly included in catalogues of AI technology. On the other hand, they have a great deal in common with it, and with the notion of complementary intelligence, and research in the field is emergent (Zhaohui et al. 2016). Cyborgisation has accordingly been defined as being within-scope of the present analysis.

3.3 'AI / ML'

Computing applications for drawing inferences from data began with hard-wired, machine-level and assembler languages (1940-1960), but made great progress with genuinely 'algorithmic programs', in languages such as ForTran (formula translator). That approach involves an implied problem that needs to be solved, and an explicit procedural solution to that problem. During the 1980s, additional means of generating inferences became mainstream, including logic programming and rule-based ('expert') systems. These embody no explicit 'problem' or 'solution'. They instead define a 'problem-domain': some form of modelling of the relevant real world is undertaken, and the model is expressed in a form that enables inferences to be drawn from it.

AI research has delivered a further technique, which accords primacy to the data rather than the model, and has the effect of obscuring the model to such an extent that no humanly-understandable rationale exists for the inferences that are drawn. The relevant branch of AI is 'machine learning' (ML), and the most common technique in use is 'artificial neural networks'. The approach dates to the 1950s, but limited progress was made until sufficiently powerful processors were readily available, from the late 1980s.

Neural nets involve a set of nodes (each of which analogous to the biological concept of a neuron), with connections or arcs among them, referred to as 'edges'. Each connection has a 'weight' associated with it. Each node performs some computation based on incoming data and may as a result adapt its internal state, including the weighting on each connection, and may pass output to one or more other nodes. A neural net has to be 'trained'. This is done by selecting a training method (or 'learning algorithm') and feeding a 'training-set' of data to the network in order to load up a set of weightings on the connections between nodes.

Unlike previous techniques for developing software, neural networking approaches do not 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 is simply a list of input variables and a list of output variables (and, in the case of 'deep' networks, intermediary variables). If a model exists, in the sense of a representation of the real world, it is implicit rather than express. The weightings imputed for each connection reflect the characteristics firstly of the training-set that was fed in, and secondly of the particular learning algorithm that was imposed on the training-set.

Although algorithms are used in the imputation of weightings on the connections within a neural net, the resulting software is not algorithmic, but rather empirical. This has led some authors to justify a-theoretical mechanisms in general, and to glorify correlation and deprecate the search for causal relationships and systemic analysis generally (Anderson 2008, Mayer-Schonberger & Cukier 2013).

AI/ML may well have the capacity to discover gems of otherwise-hidden information. However, the inferences drawn inevitably reflect any errors and biasses inherent in the implicit model, in the selection of real-world phenomena for which data was created, in the selection of training-set, and in the learning algorithms used to develop the software that delivers the inferences. Means are necessary to assess the quality of the implicit model, of the data-set, of the data-item values, of the training-set and of the learning algorithm, and the compatibility among them, and to validate the inferences both logically and empirically. Unless and until those means are found, and are routinely applied, AI/ML and neural nets must be regarded as unproven techniques that harbour considerable dangers to the interests of organisations and their stakeholders.

3.4 Intellectics

Robotics began with an emphasis on machines being enhanced with computational elements and software. However, the emphasis has been shifting. I contend that the conception now needs to be inverted, and the field regarded as computers enhanced with sensors and actuators, enabling computational processes to sense the world and act directly on it. Rather than 'machines that think', the focus needs to be on 'computers that do'. The term 'intellectics' is a useful means of encapsulating that switch in emphasis.

The term has been previously used in a related manner by Wolfgang Bibel, originally in German (1980, 1989). Bibel was referring to the combination of Artificial Intelligence, Cognitive Science and associated disciplines, using the notion of the human intellect as the integrating element. Bibel's sense of the term has gained limited currency, with only a few mentions in the literature and only a few authors citing the relevant papers. The sense in which I use the term here is rather different:

> *In the new context of intellectics, artefacts go beyond merely drawing inferences from data, in that they generate a strong impulse for an action to be taken in the real world*

I suggest the following criteria for assessing whether an artefact should be classified as falling within the field of intellectics:

> *As a threshold test, the artefact must at least communicate a recommendation to a human*

> *At a higher level, an artefact makes a decision, which will result in action unless over-ridden or countermanded by a human*
>
> *At the highest level, an artefact makes a decision, and takes action in the real world to give effect to that decision, without providing an opportunity for a human to prevent the action being taken*

The effect of implementing intellectics is to at least reduce the moderating effect of humans in the decision-loop, and even to remove that effect entirely. The emergence of intellectics is accordingly bringing into much stronger focus the legitimacy of the inferencing techniques used, and of the inferences that they are leading to. Among the major challenges involved are the difficulty and expense of establishing reliable software (in particular the size of the training-set required), the low quality of a large proportion of the data on which inferencing depends, the significance of and the approach adopted to empty cells within the data-set, and the applicability of the data-analytic technique to the data to which it is applied (Clarke 2016a, 2016c).

The earlier generations of computer-performed inferencing enabled the expression of humanly-understandable explanations. During the procedural programming era, a set of conditions resulted in an output, and the logic of the solution was express in both the software specification and the source-code. In logic-based programming, 'consequents' could be traced back to 'antecedents', and in rule-based systems, which rules 'fired' in order to deliver the output could be documented (Clarke 1991).

That situation changes substantially with AI/ML and its primary technique, neural nets. The model is at best implicit and may be only very distantly related to the real-world it is assumed to represent, the approach is empirical, it depends on a training-set, and it is not capable of generating a humanly-understandable explanation for an inference that has been drawn. The application of such inferences to decision-making, and to the performance of actions in and on the real world, raises serious questions about transparency (Burrell 2016, Knight 2017). A result of the loss of decision transparency is the undermining of organisations' accountability for their decisions and actions. In the absence of transparency, such principles are under threat as evaluation, fairness, proportionality, evidence-based decision-making, and the capacity to challenge decisions (APF 2013).

Applications of a variety of data analytics techniques are already giving rise to public disquiet, even in the case of techniques that are (at least in principle) capable of generating explanations of decision rationale. The most publicly-visible of these are systems for people-scoring, most prominently in financial credit. There are also applications in 'social credit' - although in this case to date only in the PRC (Chen & Cheung 2017). Similar techniques are also applied in social welfare contexts, sometimes with seriously problematical outcomes (e.g. Clarke 2018a). Concerns are naturally heightened where inferencing technologies are applied to prediction - particularly where the technique's effectiveness is assumed rather than carefully tested, published, and subject to challenge. Such approaches result in something approaching pre-destination, through the allocation of individual people to categories and the attribution of future behaviour, in some circumstances even behaviour of a criminal nature.

There is increasing public pressure for explanations to be provided for decisions that are adverse to the interests of individuals and of small business, especially in the context of inscrutable inferencing techniques such as neural networking. The responsibility of decision-makers to provide explanations is implied by the principles of natural justice and procedural fairness. In the EU, since mid-2018, as a consequence of Articles 13.2(f), 14.2(g) and 15.1(h) of the General Data Protection Regulation (GDPR 2018), access must be provided to "meaningful information about the logic involved", "at least in" the case of automated decisions (Selbst & Powles 2017). On the other hand, "the [European Court of Justice] has ... made clear that data protection law is not intended to ensure the accuracy of decisions and decision-making processes involving personal data, or to make these processes fully transparent [and] a new data protection right, the 'right to reasonable inferences', is needed" (Wachter & Mittelstadt 2019).

Re-conception of the field as Intellectics enables focus to be brought to bear on key issues confronting organisations that apply the outcomes of AI research. Intellectics represents a major power-shift towards large organisations and away from individuals. Substantial pushback from the public needs to be anticipated, and new regulatory obligations may be imposed on organisations. The following sections canvass the scope for these concerns to be addressed firstly by ethics, and secondly through regulatory arrangements.

---

4. Ethics

Both the dated notion of AI and the alternative conceptualisations of complementary intelligence and intellectics harbour potentials for harm. So it is important for organisations to carefully consider what factors constrain their freedom of choice and actions. The following section examines the regulatory landscape. This section first considers the extent to which ethics affects organisational applications of technology.

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. In a recent paper, Floridi (2018) has referred to 'hard ethics' - that which "may contribute to making or shaping the law" - and 'soft ethics' - which are discussed after the fact.

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're a barrier to progress.

The occasional fashion of 'business ethics' naturally inherits the vagueness of ethics generally, and provides little or no concrete guidance to 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 attempts to formulate general ethical principles that may have applicability in technology-rich contexts - including bio-medicine, surveillance and information technology. Remarkably, none of them contain any explicit reference to identifying relevant stakeholders. However, a number of norms are frequently-encountered in these sets. These include demonstrated effectiveness and benefits, justification of disbenefits, mitigation of disbenefits, proportionality of negative impacts, supervision (including safeguards, controls and audit), and recourse (including complaints and appeals channels, redress, sanctions, and enforcement powers and resources).

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

When evaluating the potential impact of ethics and CSR, it is important to appreciate the constraints on company directors. They are required by law to act in the best interests of each company of which they are a director. Attention to broad ethical questions is generally 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. The primary ways in which benefits can accrue are through compliance with regulatory requirements, and enhanced relationships with important stakeholders. Most commonly, these stakeholders will be customers, suppliers and employees, but the scope might extend to communities and economies on which the company has a degree of dependence.

Given the limited framework provided by ethics, the question arises as to the extent to which organisations are subject to legal and social mechanisms that prevent or

constrain their freedom to create technologies, and to embody them in artefacts, systems and applications.

---

5. Regulation

AI seems to have been argued by its proponents to be arriving imminently, on a cyclical basis, roughly every decade since 1956. Despite that, it appears that few regulatory requirements have been designed or modified specifically with AI in mind. One reason for this is that parliaments seldom act in advance of new technologies being deployed.

A 'precautionary principle' has been enunciated, whose strong form exists in some jurisdictions' environmental laws, along the lines of 'When human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that potential harm' (TvH 2006). More generally, however, the 'principle' is merely an ethical norm to the effect that 'If an action or policy is suspected of causing harm, and scientific consensus that it is not harmful is lacking, then the burden of proof arguably falls on those taking the action'. Where AI appears likely to be impactful on the scale that its proponents suggest, surely the precautionary principle applies, at the very least in its weak form. On the other hand, the considerable impacts of such AI technologies as automated number-plate recognition (ANPR), 'facial recognition' and drones have not been the subject even of effective after-the-fact regulatory adaptation or innovation, let alone of proactive protective measures.

A large body of theory exists relating to regulatory mechanisms (Braithwaite & Drahos 2000). Regulation takes many forms, including intrinsic and natural controls, self-control, several levels of 'soft' community controls, various kinds of 'formal' or 'hard' regulatory schemes, and regulation by infrastructure or 'code'. An overview of these categories is in Clarke & Bennett Moses (2014), and a relevant analysis is in Clarke (2014c). This section identifies a range of sources that may offer organisations, to some extent guidance, and at least insights into what society expects, and obligations that organisations might be subject to.

5.1 Intrinsic and Natural Controls

An intervention such as the adoption of AI-based technologies may be subject to intrinsic limitations, or may stimulate natural processes whose effect is to prevent the adoption occurring or continuing, or to curb or mitigate negative impacts. It is appropriate to consider these first. This is because, in the absence of such harm-limitation mechanisms (a condition referred to by economists as 'market failure'), a case exists for regulatory measures to be devised and imposed; whereas, if adequate intrinsic or natural controls exist, the costs that regulation would impose on all parties are not justifiable.

Economic factors tend to constrain adoption, commonly because of the expense involved and inadequate volume or profit-margin. This is particularly likely to be determinative where the technology is, or is perceived to be, insufficiently effective in delivering on its promise. In some circumstances, the realisation of the potential benefits of a technology may be dependent on infrastructure that is unavailable or inadequate. (For example, computing could have exploded in the third quarter of the 19th century, rather than 100 years later, had metallurgy of the day been able to support Babbage's 'difference' and 'analytical' engines). Another form of control is the opposition of players with sufficient institutional or market power. This includes the use of formal media and social media to stir up public opprobrium.

It is far from clear that any of the currently-promoted forms of AI are subject to adequate intrinsic and natural controls. The following sub-sections accordingly consider each of the various forms of regulatory intervention, beginning at the apex of the regulatory pyramid with 'hard law'.

5.2 AI-Specific Laws

In-place industrial robotics, in production-lines and warehouses, is well-established. Various publications have discussed general questions of robot regulation (e.g. Leenes & Lucivero 2014, Scherer 2016, HTR 2018a, 2018b), but fewer identify AI-specific laws. Even such vital aspects as worker safety and employer liability appear to depend not on technology-specific laws, but on generic laws, which may or may not have been adapted to reflect the characteristics of the new technologies.

In HTR (2017), South Korea is identified as having enacted the first national law relating to robotics generally: the Intelligent Robots Development Distribution Promotion Act of 2008. It is almost entirely facilitative and stimulative, and barely even aspirational in relation to regulation of robotics. There is mention of a 'Charter', "including the provisions prescribed by Presidential Decrees, such as ethics by which the developers, manufacturers, and users of intelligent robots shall abide" - but no such Charter appears to exist. A mock-up is at Akiko (2012). HTR (2018c) offers a generic regulatory specification in relation to research and technology generally, including robotics and AI.

In relation to autonomous motor vehicles, a number of jurisdictions have enacted laws. See Palmerini et al. (2014, pp.36-73), Holder et al. (2016), DMV-CA (2018), Vellinga (2017), which reviews laws in the USA at federal level, California, United Kingdom, and the Netherlands, and Maschmedt & Searle (2018), which reviews such laws in three States of Australia. Such initiatives have generally had a strong focus on economic motivations, the stimulation and facilitation of innovation, exemptions from some existing regulation, and limited new regulation or even guidance. One approach to regulation is to leverage off natural processes. For example, Schellekens (2015) argued that a requirement of obligatory insurance was a sufficient means for regulating liability for harm arising from self-driving cars. In the air, legislatures and regulators have moved very slowly in relation to the regulation of drones (Clarke & Bennett Moses 2014, Clarke 2016b).

Automated decision-making about people has been subject to French data protection law for many years. In mid-2018 this became a feature of European law generally, through the General Data Protection Regulation (GDPR) Art. 22, although doubts have been expressed about its effectiveness (Wachter et al. 2017).

On the one hand, it might be that AI-based technologies are less disruptive than they are claimed to be, and that laws need little adjustment. On the other, a mythology of 'technology neutrality' pervades law-making. Desirable as it might be for laws to encompass both existing and future artefacts and processes, genuinely disruptive technologies have features that render existing laws ambiguous and ineffective.

5.3 Generic Laws

Applications of new technologies are generally subject to existing laws. Particularly with 'breakthrough', revolutionary and disruptive technologies, existing laws are likely to be ill-fitted to the new context, because they were "designed around a socio-technical context of the relatively distant past" (Bennett Moses 2011, p.765), and without knowledge of the new form. In some cases, existing law may hinder new technologies in ways that are unhelpful to both the innovators and those affected by them. In other cases, existing law may have been framed in such a manner that it does not apply to the new form (or judicial calisthenics has to be performed in order to make it appear to apply), even though there would have been benefits if it had done so.

Applications of AI will generally be subject to the various forms of commercial law, particularly contractual obligations including express and implied terms, consumer rights laws, and copyright and patent laws. In some contexts (such as robotics, cyborg artefacts, and AI software embedded in devices), product liability laws may apply. Other laws that assign risk to innovators may also apply, such as the tort of negligence, as may laws of general applicability such as human rights law, anti-discrimination law and data protection law. The obligations that the corporations law assigns to company directors are also relevant. Further sources of regulatory impact are likely to be the laws relating to the various industry sectors within which AI is applied, such as road transport law, workplace and employment law, and health law.

Particularly in common law jurisdictions, there is likely to be a great deal of uncertainty about the way in which laws will be applied by tribunals and courts if any particular dispute reaches them. This acts to some extent as a deterrent against innovation, and can considerably increase the costs incurred by proponents, and delay deployment. From the viewpoint of people who perceive themselves to be negatively affected by the innovation, on the other hand, channels for combatting those threats

may be inaccessible, expensive, slow and even entirely ineffectual.

5.4 Co-Regulation (Enforceable Codes)

Parliaments struggle to understand and cope with new technologies. An approach to regulation that once appeared to offer promise is co-regulation. Under this arrangement, a parliament establishes a legal framework, including authority, obligations, sanctions and enforcement mechanisms, but without expressing the obligations at a detailed level. This is achieved through consultative processes among advocates for the various stakeholders. The result is an enforceable Code, which articulates general principles expressed in the relevant legislation.

Unfortunately, few instances of effective co-regulation exist, because such processes typically exclude less powerful stakeholders. In any case, there are few signs of parliaments being aware of the opportunity, and of its applicability to Intellectics. In Australia, for example, Enforceable Codes exist that are administered by the Australian Communications and Media Authority (ACMA) in respect of TV and radio broadcasting, and telecommunications, and by the Australian Prudential Regulation Authority (APRA) in respect of banking services. These arrangements succeed both in facilitating business and government activities and in offering a veneer of regulation; but they fail to exercise control over behaviour that the public regards as inappropriate, and hence they have little public credibility.

5.5 Guidance by Regulatory and Oversight Agencies

It is common for parliaments to designate a specialist government agency or parliamentary appointee either to exercise loose oversight over a contested set of activities, or to exercise powers and resources in order to enforce laws or Codes. An important function of either kind of organisation is to provide guidance to both the regulatees and the parties that the scheme is intended to protect. In very few instances, however, does it appear that AI lies within the scope of an existing agency or appointee. Some exceptions may exist, for example in relation to the public safety aspects of drones and self-driving motor vehicles.

As a result, in most jurisdictions, limited guidance appears to exist. For example, six decades after the AI era was launched, the EU has gone no further than a preliminary statement (EC 2018) and a discussion document issued by the Data Protection Supervisor (EDPS 2016). Similarly, the UK Data Protection Commissioner has only reached the stage of issuing a discussion paper (ICO 2017). The current US Administration's policy is entirely stimulative in nature, and mentions regulation solely as a barrier to economic objectives (WH 2018).

5.6 Industry Self-Regulation (Unenforceable Codes)

Corporations club together for various reasons, some of which can be to the detriment of other parties, such as collusion on bidding and pricing. The activities of industry associations can, however, deliver benefits for others, as well as for their members. In particular, collaborative approaches to infrastructure can improve services and reduce costs for the sector's customers.

It could also be argued that, if norms are promulgated by the more responsible corporations in an industry sector, then misbehaviour by the industry's 'cowboys' would be highlighted. In practice, however, the effect of Industry Codes on corporate behaviour is seldom significant. Few such Codes are sufficiently stringent to protect the interests of other parties, and the absence of enforcement undermines the endeavour. The more marginal kinds of suppliers ignore them, and responsible corporations feel the pinch of competition and reduce their commitment to them. As a result, such Codes act as camouflage, obscuring the absence of safeguards and thereby holding off actual regulatory measures. In the AI field, examples of industry coalitions eagerly pre-countering the threat of regulation include FLI (2017), ITIC (2017), and PoAI (2018).

A more valuable role is played by industry standards. HTR (2017) lists industry standards issued by the International Standards Organisation (ISO) in the AI arena. A considerable proportion of industry standards focus on inter-operability, and on business processes intended to achieve quality assurance. Public safety is also an area of strength, particularly in the field commonly referred to as 'safety-critical systems' (e.g. Martins & Gorschek 2016). Hence some of the physical threats embodied in AI-based systems are able to be avoided, mitigated and managed through the development and application of industry standards; but threats to economic and social interests are seldom addressed.

A role can also be played by professional associations, because these generally balance public needs against self-interest somewhat better than industry associations. Their impact is, however, far less pronounced than that of industry associations. Moreover, the intiatives to date of the two largest bodies are underwhelming, with ACM (2017) using weak forms such as "should" and "are encouraged to", and IEEE (2017) offering lengthy prose but unduly vague and qualified principles. Neither has to date provided the guidance needed by professionals, managers and executives.

5.7 Organisational Self-Regulation

It was noted above that Directors of corporations are required by law to pursue the interests of the corporation ahead of all other interests. It is therefore unsurprising, and even to be expected, that organisational self-regulation is almost always ineffectual from the viewpoint of the supposed beneficiaries, and often not even effective at protecting the organisation itself from bad publicity. Recent offerings by major corporations include IBM (Rayome 2017), Google (Pichai 2018) and MS (2018). For an indication of the scepticism with which such documents are met, see Newcomer (2018).

5.8 Vague Prescriptions

A range of, in most cases fairly vague, principles, have been proposed by a diverse array of organisations. Examples include the European Greens Alliance (GEFA 2016), a British Standard BS 8611 (BS 2016), the UNI Global Union (UGU 2017), the Japanese government (Hirano 2017), a House of Lords Committee (HOL 2018), as interpreted by a World Economic Forum document (Smith 2018), and the French Parliament (Villani 2018).

Although there are commonalities among these formulations, there is also a lot of diversity, and few of them offer usable advice on how to ensure that Intellectics is applied in a responsible manner. The next section draws on the sources identified above, in order to offer practical advice. It places the ideas within a conventional framework, but extends that framework in order to address the needs of all stakeholders rather than just the corporation itself.

5.9 Regulation by 'West Coast Code'

One further regulatory element requires consideration. Lessig (1999) popularised the notion of behaviour in socio-technical systems being subject not only to formal law ('East Coast Code'), but also to constraints that exist within computer and network architecture and infrastructure, i.e. standards, protocols, hardware and software ('West Coast Code').

A relevant form that 'West Coast Code' could take is the embedment in robots of something resembling 'laws of robotics'. The notion dates to an Asimov short story, 'Runaround', first published in 1942; but many commentators on robotics cling to it. For example. Devlin (2016) quotes a professor of robotics as perceiving that the British Standard Institute's guidance on ethical design of robots (BS 2016) represents "the first step towards embedding ethical values into robotics and AI". On the other hand, a study of Asimov's robot fiction showed that he had comprehensively demonstrated the futility of the idea (Clarke 1993). A recent expression of the reason why the approach is doomed is that "You cannot construct an algorithm that will reliably decide whether or not any algorithm is ethical" (Castell 2018, p.743).

6. A Practical Approach

Ethical analyses offer little assistance, and regulatory frameworks are lacking. It might seem attractive to business enterprises to face few legal obligations and hence to be subject to limited compliance risk exposure. On the other hand, the absence of regulation heightens many other business risks. At least some competitors inevitably exhibit 'cowboy' behaviour, and there are always individuals and groups within each organisation who can be tempted by the promise that AI appears to offer. As a result, there are substantial direct and indirect threats to the organisation's reputation. It is therefore in each organisation's own self-interest for a modicum of regulation to exist, in order to provide a protective shield against media exposés and public backlash.

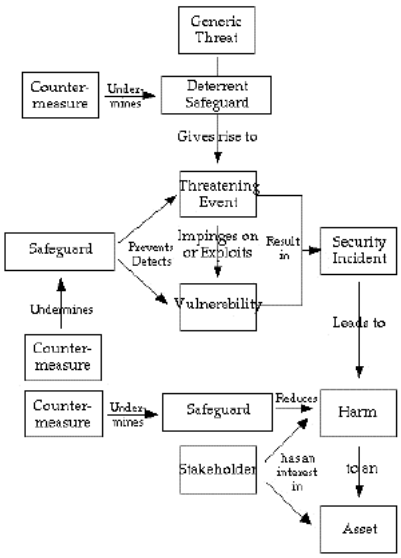
This section offers guidance to organisations. It assumes that organisations evaluating AI apply conventional environmental scanning and marketing techniques in order to identify opportunities, and a conventional business case approach to estimating the strategic, market-share, revenue, cost and profit benefits that the opportunities appear to offer them. The focus here is on how the downsides can be identified, evaluated and managed.

Familiar, practical approaches to assessing and managing risks are applicable. However, I contend that the conventional framework must be extended to include an important element that is commonly lacking in business approaches to risk. That missing ingredient is stakeholder analysis. Risk assessment and management needs to be performed not only from the business perspective, but also from the perspectives of other stakeholders.

6.1 Corporate Risk Assessment

There are many sources of guidance in relation to risk assessment and management. The techniques are well-developed in the context of 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 Figure 1.

Figure 1: The Conventional Risk Model



Existing corporate practice approaches this model from the perspective of the organisation itself. This gives rise to conventional risk assessment and risk management processes outlined in Table 2. Relevant assets are identified, and an analysis 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 incidents. Existing safeguards are taken into account, in order to guide the development of a strategy and plan to refine and extend the safeguards and thereby provide a degree of protection that is judged to suitably balance modest actual costs against much higher contingent costs.

Table 2: The Risk Assessment and Risk Management Processes

Analyse / Perform Risk Assessment
(1) Define the Objectives and Constraints
(2) Identify the relevant Stakeholders, Assets, Values and categories of Harm
(3) Analyse Threats and Vulnerabilities
(4) Identify existing Safeguards
(5) Identify and Prioritise the Residual Risks
Design / Initiate Risk Management
(1) Identify alternative Safeguards
(2) Evaluate the alternatives against the Objectives and Constraints
(3) Select a Design (or adapt / refine the alternatives to achieve an acceptable Design)

Do / Perform Risk Management
(1) Plan the implementation
(2) Implement
(3) Review the implementation

6.2 Stakeholder Risk Assessment

The notion of 'stakeholders' was introduced as a means of juxtaposing the interests of other parties against those of the corporation's shareholders (Freeman & Reed 1983). Many stakeholders are participants in relevant processes, in such roles as employees, customers and suppliers. Where the organisation's computing services extend beyond its boundaries, any and all of those primary categories of stakeholder may be users of the organisation's information systems.

However, the categories of stakeholders are broader than this, comprising not only "participants in the information systems development process" 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 & Whitley 1997, p.3). The term 'usees' is a usefully descriptive term for these once-removed stakeholders (Clarke 1992, Fischer-Hübner & Lindskog 2001, Baumer 2015).

My first proposition for extension beyond conventional corporate risk assessment is that the responsible application of AI is only possible if stakeholder analysis is undertaken in order to identify the categories of entities that are or may be affected by the particular project (Clarkson 1995). There is a natural tendency 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 the responsible application of AI depends on risk assessment processes being conducted from the perspectives of the various stakeholders, to complement that undertaken from the perspective of the corporation. Conceivably, such assessments could be conducted by the stakeholders independently, and fed into the organisation. In practice, the asymmetry of information, resources and power is such that the outputs from independent. and therefore uncoordinated, activities are unlikely to gain acceptance. The responsibility lies with the sponsor of an initiative to drive the studies, engage effectively with the other parties, and reflect their input in the project design criteria and features.

The risk assessment process outlined in Table 2 above is generally applicable. However, my third proposition is that risk assessment processes that reflect the interests of stakeholders needs to be broader than that commonly undertaken within organisations. Relevant techniques include privacy impact assessment (Clarke 2009, Wright & De Hert 2012), social impact assessment (Becker & Vanclay 2003), and technology assessment (OTA 1977). For an example of impact assessment applied to the specific category of person-carrier robots, see Villaronga & Roig (2017). The most practical approach may be, however, to adapt the organisation's existing process in order to encompass whichever aspects of such broader techniques are relevant to the stakeholders whose needs are being addressed.

6.3 Comprehensive Risk Management

The results of the two or more risk assessment processes outlined above deliver the information that the organisation needs. They enable the development of a strategy and plan whereby existing safeguards can be 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. A framework is presented in Table 3 that in my experience is more understandable by practitioners and more readily usable as a basis for identifying possible safeguards.

Table 3: 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, publicity for prosecutions, substantial fines, gaol-time
- **Prevention**
  e.g. surge protectors and backup power sources; quality equipment, media and software; physical and logical access control; staff training, assigned responsibilities and measures to sustain morale; staff termination procedures
- **Redundancy**
  e.g. duplicated equipment and communication paths; multiple, parallel evaluations with cross-checking of results

Reactive Strategies

- **Detection**
  e.g. fire and smoke detectors, logging, log-analysis, exception reporting
- **Reduction / Mitigation**
  e.g. fire-suppression technologies, fire-warden training, suspension of processing when unexpected harm arises, pre-arranged contingent measures to compensate for harm
- **Recovery**
  e.g. investment in resources, procedures/documentation, staff training, and duplication including 'hot-sites' and 'warm-sites'
- **Insurance**
  e.g. mutual arrangements with other organisations, 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. siting a nuclear energy company's headquarters adjacent to the power plant, on the grounds that, if it goes, then the organisation and its employees should go with it

---

Existing techniques are strongly oriented towards protection against risks as perceived by the organisation. Risks to other stakeholders are commonly treated as, at best, a second-order consideration, and at worst as if they were out-of-scope. All risk management work involves the exercise of a considerable amount of imagination. That characteristic needs to be underlined even more strongly in the case of the comprehensive, multi-stakeholder approach that I am contending is necessary in the case of AI-based systems.

This section has suggested customisation of existing, generic techniques in order to address the context of AI-based systems. The following section presents more specific proposals.

---

7. Towards Operational Principles

This section presents a set of Principles for AI. The purpose of doing so is to provide organisations and individuals with guidance as to how they can fulfil their responsibilities in relation to AI and AI-based activities. Because of the broad scope of the AI notion, the considerable diversity among its various forms, and the changes in those forms over time, the Principles proposed below are still somewhat abstract. The intention is to express them in as practically useful a manner as can reasonably be achieved. At the very least, they should provide a firm base for the expression of operational guidance for each specific form of AI.

The Principles in part emerge from the analysis presented in this Working Paper, and in part represent a consolidation of ideas from a suite of previously-published sets of principles. 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 were available in the english language were used. This resulted in a strong bias within the suite towards documents that originated in countries whose primary language(s) is or include English. Of the individual documents, 8 are formulations of 'ethical principles and IT'. Extracts and citations are provided at Clarke (2018c). The other 16 claim to provide principles or guidance specifically in relation to AI. Extracts and citations are at Clarke (2018d).

In s.2.3 and Table 1 above, distinctions were drawn among the 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. In only a few of the 24 documents in the suite were such distinctions evident, however, 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. Overall, however, most of the source documents were remarkably sparse, and there was far less consensus that might have been expected 60 years after AI was first heralded. For example, only 1 document encompassed cyborgisation (GEFA 2016); only 2 documents referred to the precautionary principle (CLA-EP 2016, GEFA 2016), and only 5 stipulated the conduct of impact assessments. One striking statistic is that only 3 of the c. 50 Principles were 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 just one form of AI, such as robotics or machine-learning. Documents prepared by corporations, industry associations, and even professional associations and joint associations tended to adopt the perspective of producer roles, with the interests of other stakeholders often relegated to a secondary consideration. For example, 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 (FLI 2017 at 3). As a result, transparency is constrained to a small sub-set of circumstances (at 6), '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), and "strict safety and control measuers" are limited to a small sub-set of AI systems (at 22). 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 regards decision opaqueness 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 favourably reported in Smith (2018), it appears to have adopted the perspective of the AI industry.

The process of developing the set commenced with themes that derived from the analysis reported on in the earlier sections of this Working Paper. The previously-published sets of principles were then inspected. Detailed propositions within each set were 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 in order to sustain coherence and limit the extent to which duplications arise.

The Principles 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. The full set of Principles, comprising about 50 elements, is in Appendix 1. In order to make them more digestible, Table 4 presents the 10 over-arching themes.

Some of the items that appear in source documents appear incapable of being operationalised. For example, 'human dignity', 'fairness' and 'justice' are vague abstractions that need to be unpacked into more specific concepts. In addition, some items fall outside the scope of the present work. The items that have been excluded from the set in Table 4 are listed in Appendix 2.

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. 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 be legal requirements. They are expressions of moral obligations; but no authority exists that can impose such obligations. In addition, all are contestable, and in different

circumstances any of them may be in conflict with other legal or moral obligations, and with various interests of various stakeholders. They represent guidance to organisations involved in AI as to the expectations of courts, regulatory agencies, oversight agencies, competitors and stakeholders. They are intended to be taken into account as organisations undertake risk assessment and risk management, as outlined in s.6 above.

Table 4: Principles For A.I. Technologies, Artefacts, Systems and Applications

The following Principles are intended to be applied by the entities responsible for all phases 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 applying AI bear legal and moral responsibility to demonstrate the benefits, to be proactive in relation to disbenefits, and to involve stakeholders in the process.

2. Complement Humans

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

4. Ensure Human Safety and Wellbeing

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

5. Ensure Consistency with Human Values and Human Rights

AI is capable of having substantial negative impacts on a wide range of civil and political rights.

6. Embed Quality Assurance

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

7. Deliver Transparency and Auditability

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

8. Exhibit Robustness and Resilience

AI-based systems and associated data must be subject to safeguards commensurate with the significance of their benefits, sensitivity and potential to cause harm to stakeholders.

9. Ensure Accountability for Legal and Moral Obligations

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

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

All entities involved in 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.

The Principles in Table 4 are intentionally framed and phrased in an abstract manner, in an endeavour to achieve applicability to at least the currently mainstream forms of AI discussed earlier - robotics, particularly remote-controlled and self-driving vehicles; cyborgs who incorporate computational capabilities; and AI/ML/neural-networking applications. More broadly, the intention is that they be applicable to what I proposed above as the appropriate conceptualisation of the field - Intellectics.

These Principles are capable of being further articulated into much more specific guidance in respect of each particular category of AI. For example, in a companion project, I have proposed 'Guidelines for Responsible Data Analytics' (Clarke 2018b). These provide more detailed guidance for the conduct of all forms of data analytics projects, including those that apply AI/ML/neural-networking approaches. Areas addressed by the Data Analytics guidelines 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.

---

8. Conclusions

This paper has proposed that the unserviceable notion of AI should be replaced by the notion of 'complementary intelligence', and that the notion of robotics ('machines that think') is now much less useful than that of 'intellectics' ('computers that do').

The techniques and technologies that emerge from research laboratories offer potential but harbour considerable threats to organisations, and to those organisations' stakeholders. Sources of guidance have been sought, whereby organisations in both the private and public sectors can evaluate the appropriateness of various such technologies to their own operations. Neither ethical analysis nor regulatory schemes deliver what organisations need. The paper concludes that adapted forms of risk assessment and risk management processes can fill the void, and that principles specific to AI can be formulated.

The propositions in this paper need to be workshopped with colleagues in the academic and consultancy worlds. The abstract Principles need to be articulated into more specific expressions that are directly relevant to particular categories of technology, artefacts, systems and applications. The resulting guidance then needs to be exposed to relevant professional executives and managers, reviewed by internal auditors, government relations executives and corporate counsel, and pilot-tested in realistic settings.

Appendix 1: 50 Principles for Responsible AI Technologies, Artefacts, Systems and Applications

A PDF version of this Appendix is available

The following Principles are intended to be applied by the entities responsible for all phases of AI research, invention, innovation, dissemination and application. The cross-references are to the 'Ethical Principles and IT' sources (Clarke 2018b - E) and 'Principles for AI' sources (Clarke 2018c - P).

1. Evaluate Positive and Negative Impacts

1.1 Conceive and design only after ensuring adequate understanding of purposes and contexts
(E4.3, P5.3, P6.21, P7.1, P15.7)

1.2 Justify objectives
(E3.25)

1.3 Demonstrate the achievability of postulated benefits
(Not found in any of the documents, but a logical pre-requisite)

1.4 Conduct impact assessment
(E7.1, P3.12, P4.1, P4.2, P6.21, P11.8)

1.5 Publish sufficient information to stakeholders to enable them to conduct impact assessment
(E7.3, P3.7, P4.1, P8.3, P8.4, P8.7)

1.6 Conduct consultation with stakeholders and enable their participation
(E5.6, E7.2, P3.7, P8.6, P8.7)

1.7 Justify negative impacts on individuals ('proportionality')
(E3.21, E7.4, E7.5)

1.8 Consider alternative, less harmful ways of achieving the same objectives
(E3.22)

2. Complement Humans

2.1 Design as an aid, for augmentation, collaboration and inter-operability
(P4.5, P5.1, P9.1, P9.8 .P14.2, P14.4)

2.2 Avoid design for replacement of people by independent devices, except in circumstances in which artefacts are demonstrably more capable than people, and even then ensuring that the result is complementary to human capabilities
(P5.1)

3. Ensure Human Control

3.1 Ensure human control over AI-based artefacts and systems
(E4.2, E6.1, E6.8, E6.19, P1.4, P2.1, P4.2, P5.2, P6.16, P8.4, P9.3, P12.1, P13.5, P15.4)

3.2 In particular, ensure control over autonomous behaviour of AI-based artefacts and systems
(E8.1, P8.4, P10.2, P11.4)

3.3 Respect each person's autonomy, freedom of choice and self-determination
(E2.1, E5.3, P3.3, P9.7, P11.3)

3.4 Ensure human review of inferences and decisions prior to acting on them
(E3.11)

3.5 Respect people's expectations in relation to personal data protections (E5.6), including:
* awareness of data-usage (E3.6)
* consent (E3.7, E3.28, E5.3, P3.11, P4.6)
* data minimisation (E3.9)
* public visibility and consultation (E3.10, E7.2), and
* relationship of data-usage to the data's original purpose (E3.27)

3.6 Avoid deception of humans
(E4.4, E6.20, P2.5)

3.7 Avoid services being conditional on the acceptance of AI-based artefacts and systems
(P4.5)

4. Ensure Human Safety and Wellbeing

4.1 Ensure people's physical health and safety ('nonmaleficence')
(E2.2, E3.1, E4.1, E4.3, E5.4, E6.2, E6.9, E6.13, E6.14, E6.18, P1.2, P1.3, P2.1, P3.2, P3.6, P3.9, P3.12, P4.3, P4.9, P6.6, P8.4, P9.4, P10.2, P11.4, P13.5, P14.1, P15.3)

4.2 Ensure people's psychological safety (E3.1, E6.9, E6.13), by avoiding negative effects on any individual's mental health, inclusion in society, worth, standing in comparison with other people, or emotional state (E5.4, E6.3)

4.3 Ensure people's wellbeing ('beneficence')
(E2.3, E3.20, E5.5, P3.1, P3.4, P6.1, P6.14, P6.15, P8.6, P11.6, P12.2, P13.1, P15.1, P16.4)

4.4 Mitigate negative consequences
(E3.24, E7.6, E6.21, E10.4)

4.5 Avoid violation of trust
(E3.3)

4.6 Avoid the manipulation of vulnerable people (E4.4, P4.5, P4.9), including taking advantage of individuals' tendency to addiction, e.g. to gambling (E6.3)

5. Ensure Consistency with Human Values and Human Rights

5.1 Ensure compliance with human rights laws
(E4.2, P3.5, P3.9, P4.3)

5.2 Be just / fair / impartial and treat individuals equally
(E2.4, E3.16, E3.29, P3.4)

5.3 Avoid unfair discrimination and bias, not only where it is legally procribed but also where it is publicly unacceptable
(ICCPR Arts. 2.1, 3, 26 and 27, E3.16, P3.4, P4.5, P11.5, P15.2, P16.1)

5.4 Avoid restrictions on freedom of movement
(ICCPR 12, P6.13)

5.5 Avoid interference with privacy, family, home or reputation
(ICCPR 17, E5.6, P3.11, P6.12, P8.4, P9.6, P13.3, P15.5)

5.6 Avoid interference with the rights of freedom of information, opinion and expression
(ICCPR 19, P4.6)

5.7 Avoid interference with the right of freedom of assembly
(ICCPR 21, P6.13)

5.8 Avoid interference with the right of freedom of association
(ICCPR 22, P6.13)

5.9 Avoid interference with the rights to participation in public affairs and access to public service
(ICCPR 25, P6.13)

6. Embed Quality Assurance

6.1 Invest in quality assurance
(E6.2, P4.2, P15.6)

6.2 Ensure effective, efficient and adaptive performance of intended functions
(E6.11, P1.6)

6.3 Ensure security safeguards against inappropriate modification to and deletion of sensitive data
(E3.15)

6.4 Ensure justification of the use of sensitive data
(E3.26, E7.4)

6.5 Ensure data quality and data relevance
(P10.3, P11.2)

6.6 Deal fairly with people (faithfulness, fidelity)
(E2.5, E3.2)

6.7 Avoid invalid and unvalidated techniques
(E3.5, P7.7)

6.8 Test for result validity
(E3.5, P7.7. P9.2)

6.9 Impose controls in order to ensure that safeguards are operative and effective
(E7.7, P10.2)

6.10 Conduct audits of safeguards and controls (E7.8, P9.3)

7. Deliver Transparency and Auditability

7.1 Ensure that the fact that the process is AI-based is transparent to all stakeholders
(E4.4, P4.8)

7.2 Ensure that the means whereby inferences are drawn, decisions made and actions are taken are logged and can be reconstructed
(E6.6, P2.4, P4.8, P5.2, P6.7, P7.4, P7.6, P8.2. P9.2, P11.1, P11.2, P13.2, P16.2, P16.3)

7.3 Ensure people are aware of inferences and how they were reached
(E3.12, P2.4)

8. Exhibit Robustness and Resilience

8.1 Provide and sustain appropriate security safeguards against compromise of intended functions arising from both passive threats and active attacks (E4.3, E6.11, P1.4, P1.5, P4.9, P6.6, P8.4, P9.5)

8.2 Provide and sustain appropriate security safeguards against inappropriate access to sensitive data arising from both passive threats and active attacks (E3.15, E6.5, E6.10, P3.11, P9.6)

8.3 Conduct audits of justification, proportionality, transparency, mitigation measures and controls (E7.8, E8.4)

8.4 Ensure resilience, in the sense of prompt and effective recovery from incidents

9. Ensure Accountability for Legal and Moral Obligations

9.1 Ensure that the responsible entity is apparent or can be readily discovered by any party (E4.5, E6.4, P2.3, P3.8, P4.7, P8.5, P12.3)

9.2 Ensure that effective remedies exist, in the form of complaints processes, appeals processes, and redress where harmful errors have occurred (ICCPR 2.3, E3.13, E3.14, E7.7, P3.11, P4.7, P7.2, P8.7, P9.9, P10.5, P11.9, P16.3)

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

10.1 Ensure that complaints, appeals and redress processes operate effectively (ICCPR 2.3, E7.7)

10.2 Comply with external complaints, appeals and redress processes and outcomes (ICCPR 14), including, in particular, provision of timely, accurate and complete information relevant to cases

---

Appendix 2: Omitted Elements

The following elements within the sources have not been reflected in Table 4 and Appendix 1.

- Environmental Sustainability (E5.5, E6.7, P4.4, P11.3)
- "A user must not use a robot to commit an illegal act" (E6.12)
- "[Do not] deliberately damage or destroy a robot" (E6.15)
- "[Do not,] through gross negligence ... allow a robot to come to harm" (E6.16)
- "It is a lesser but nonetheless serious offence to treat a robot in a way which may be construed as deliberately and inordinately abusive" (E6.17)
- "[Respect a robot's] right to exist without fear of injury or death" (E6.21)
- "[Respect a robot's] right to live an existence free from systematic abuse (E6.22)
- "Fund research in particular with regards to the ethical and legal effects of artificial intelligence" (P4.10, P6.2)
- Asimov's Meta-Law (P1.1)
- Asimov's Procreation Law (P1.7)
- Ensure reversability of actions (P3.10)
- Respect and improve social processes, and avoid subverting them (P6.17)
- Public Empowerment - The public's ability to understand AI-enabled services, and how they work, is key to ensuring trust in the technology - 'Algorithmic Literacy' must be a basic skill ... (P8.3)
- Equip AI systems with an 'Ethical Black Box' that contains clear data and information on the ethical considerations built into said system (P11.2)
- Secure a Just Transition as workers are displaced (P11.7)
- Establish governance mechanisms [i.e. long-term impact assessment and management] (P11.8)

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

Akiko (2012) 'South Korean Robot Ethics Charter 2012' Akiko's Blog, 2012, at https://akikok012um1.wordpress.com/south-korean-robot-ethics-charter-2012/

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

Anderson C. (2008) 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete' Wired Magazine 16:07, 23 June 2008, at http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

APF (2013) 'Meta-Principles for Privacy Protection' Australian Privacy Foundation, March 2013, at https://privacy.org.au/policies/meta-principles/

Baumer E.P.S. (2015) 'Usees' Proc. 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15), April 2015

Becker H. & Vanclay F. (2003) 'The International Handbook of Social Impact Assessment' Cheltenham: Edward Elgar, 2003

Bennett Moses L. (2011) 'Agents of Change: How the Law Copes with Technological Change' Griffith Law Review 20, 4 (2011) 764-794, at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2000428

Bibel W. (1980) "Intellektik' statt 'KI' -- Ein ernstgemeinter Vorschlag' Rundbrief der Fachgruppe Künstliche Intelligenz in der Gesellschaft für Informatik, 22, 15-16 December 1980

Bibel W. (1989) 'The Technological Change of Reality: Opportunities and Dangers' AI & Society 3, 2 (April 1989) 117-132

Braithwaite B. & Drahos P. (2000) `Global Business Regulation' Cambridge University Press, 2000

BS (2016) 'Robots and robotic devices - Guide to the ethical design and application of robots and robotic systems' BS 8611, British Standards Institute, April 2016

Burrell J. (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms' Big Data & Society 3, 1 (January-June 2016) 1-12

Calo R. (2017) 'Artificial Intelligence Policy: A Primer and Roadmap' UC Davis L. Rev. 51 (2017) 399-404

Castell S. (2018) 'The future decisions of RoboJudge HHJ Arthur Ian Blockchain: Dread, delight or derision?' Computer Law & Security Review 34, 4 (Jul-Aug 2018) 739-753

Chen Y. & Cheung A.S.Y. (2017). 'The Transparent Self Under Big Data Profiling: Privacy and Chinese Legislation on the Social Credit System, The Journal of Comparative Law 12, 2 (June 2017) 356-378, at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2992537

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//NONSGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN

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. (1991) 'A Contingency Approach to the Application Software Generations' Database 22, 3 (Summer 1991) 23-34, PrePrint at http://www.rogerclarke.com/SOS/SwareGenns.html

Clarke R. (1992) 'Extra-Organisational Systems: A Challenge to the Software Engineering Paradigm' Proc. IFIP World Congress, Madrid, September 1992, at http://www.rogerclarke.com/SOS/PaperExtraOrgSys.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. (2009) 'Privacy Impact Assessment: Its Origins and Development' Computer Law & Security Review 25, 2 (April 2009) 123-135, PrePrint at http://www.rogerclarke.com/DV/PIAHist-08.html

Clarke R. (2011) 'Cyborg Rights' IEEE Technology and Society 30, 3 (Fall 2011) 49-57, at http://www.rogerclarke.com/SOS/CyRts-1102.html

Clarke R. (2014a) 'Understanding the Drone Epidemic' Computer Law & Security Review 30, 3 (June 2014) 230-246, PrePrint at http://www.rogerclarke.com/SOS/Drones-E.html

Clarke R. (2014b) '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. (2014c) 'The Regulation of of the Impact of Civilian Drones on Behavioural Privacy' Computer Law & Security Review 30, 3 (June 2014) 286-305, PrePrint at http://www.rogerclarke.com/SOS/Drones-BP.html

Clarke R. (2015) 'The Prospects of Easier Security for SMEs and Consumers' Computer Law & Security Review 31, 4 (August 2015) 538-552, PrePrint at http://www.rogerclarke.com/EC/SSACS.html

Clarke R. (2016a) 'Big Data, Big Risks' Information Systems Journal 26, 1 (January 2016) 77-90, PrePrint at http://www.rogerclarke.com/EC/BDBR.html

Clarke R. (2016) 'Appropriate Regulatory Responses to the Drone Epidemic' Computer Law & Security Review 32, 1 (Jan-Feb 2016) 152-155, PrePrint at http://www.rogerclarke.com/SOS/Drones-PAR.html

Clarke R. (2016c) 'Quality Assurance for Security Applications of Big Data' Proc. EISIC'16, Uppsala, 17-19 August 2016, PrePrint at http://www.rogerclarke.com/EC/BDQAS.html

Clarke R. (2018a) 'Centrelink's Big Data 'Robo-Debt' Fiasco of 2016-17' Xamax Consultancy Pty Ltd, January 2018, at http://www.rogerclarke.com/DV/CRD17.html

Clarke R. (2018b) '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. (2018c) 'Ethical Principles and Information Technology' Xamax Consultancy Pty Ltd, rev. September 2018, at http://www.rogerclarke.com/EC/GAIE.html

Clarke R. (2018d) 'Principles for AI: A 2017-18 SourceBook' Xamax Consultancy Pty Ltd, rev. September 2018, at http://www.rogerclarke.com/EC/GAI.html

Clarke R. & Bennett Moses L. (2014) 'The Regulation of Civilian Drones' Impacts on Public Safety' Computer Law & Security Review 30, 3 (June 2014) 263-285, PrePrint at http://www.rogerclarke.com/SOS/Drones-PS.html

Clarkson M.B.E. (1995) 'A Stakeholder Framework for Analyzing and Evaluating Corporate Social Performance' The Academy of Management Review 20, 1 (Jan.1995) 92-117 , at https://www.researchgate.net/profile/Mei_Peng_Low/post/Whats_corporate_social_performance_related_to_CSR/attachment/59d6567879197b80779ad3f2/AS%3A5304080

Devlin H. (2016). 'Do no harm, don't discriminate: official guidance issued on robot ethics' The Guardian, 18 Sep 2016, at https://www.theguardian.com/technology/2016/sep/18/official-guidance-robot-ethics-british-standards-institute

DMV-CA (2018) 'Autonomous Vehicles in California' Califiornian Department of Motor Vehicles, February 2018, at https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/bkgd

EC (2018) 'Statement on Artificial Intelligence, Robotics and `Autonomous' Systems' European Group on Ethics in Science and New Technologies' European Commission, March 2018, at http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf

EDPS (2016) 'Artificial Intelligence, Robotics, Privacy and Data Protection' European Data Protection Supervisor, October 2016, at https://edps.europa.eu/sites/edp/files/publication/16-10-19_marrakesh_ai_paper_en.pdf

ENISA (2016) 'Risk Management:Implementation principles and Inventories for Risk Management/Risk Assessment methods and tools' European Union Agency for

Network and Information Security, June 2016, at https://www.enisa.europa.eu/publications/risk-management-principles-and-inventories-for-risk-management-risk-assessment-methods-and-tools

Fieser J. (1995) 'Ethics' Internet Encyclopaedia of Philosophy, 1995, at https://www.iep.utm.edu/ethics/

Firesmith D. (2004) 'Specifying Reusable Security Requirements' Journal of Object Technology 3, 1 (Jan-Feb 2004) 61-75, at http://www.jot.fm/issues/issue_2004_01/column6

Fischer-Hübner S. & Lindskog H. (2001) 'Teaching Privacy-Enhancing Technologies' Proc. IFIP WG 11.8 2nd World Conference on Information Security Education, Perth, 2001, at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.24.3950&rep=rep1&type=pdf

FLI (2017) 'Asilomar AI Principles' Future of Life Institute, January 2017, at https://futureoflife.org/ai-principles/?cn-reloaded=1

Floridi L. (2018) 'Soft Ethics: Its Application to the General Data Protection Regulation and Its Dual Advantage' Philosophy & Technology 31, 2 (June 2018) 163-167, at https://link.springer.com/article/10.1007/s13347-018-0315-5

Freeman R.E. & Reed D.L. (1983) 'Stockholders and Stakeholders: A New Perspective on Corporate Governance' California Management Review 25:, 3 (1983) 88-106, at https://www.researchgate.net/profile/R_Freeman/publication/238325277_Stockholders_and_Stakeholders_A_New_Perspective_on_Corporate_Governance/links/5893a4b2a and-Stakeholders-A-New-Perspective-on-Corporate-Governance.pdf

GDPR (2018) 'General Data Protection Regulation' Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, at http://www.privacy-regulation.eu/en/index.htm

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

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

Holder C., Khurana V., Harrison F. & Jacobs L. (2016) 'Robotics and law: Key legal and regulatory implications of the robotics age (Part I of II)' Computer Law & Security Review 32, 3 (May-Jun 2016) 383-402

HTR (2017) 'Robots: no regulatory race against the machine yet' The Regulatory Institute, April 2017, at http://www.howtoregulate.org/robots-regulators-active/#more-230

HTR (2018a) 'Report on Artificial Intelligence: Part I - the existing regulatory landscape' The Regulatory Institute, May 2018, at http://www.howtoregulate.org/artificial_intelligence/

HTR (2018b) 'Report on Artificial Intelligence: Part II - outline of future regulation of AI' The Regulatory Institute, June 2018, at http://www.howtoregulate.org/aipart2/#more-327

HTR (2018c) 'Research and Technology Risks: Part IV - A Prototype Regulation' The Regulatory Institute, March 2018, at http://www.howtoregulate.org/prototype-regulation-research-technology/#more-298

ICO (2017) 'Big data, artificial intelligence, machine learning and data protection' UK Information Commissioner's Office, Discussion Paper v.2.2, September 2017, at https://ico.org.uk/for-organisations/guide-to-data-protection/big-data/

IEEE (2017) 'Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS)' IEEE, Version 2, December 2017, at http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

ISM (2017) 'Information Security Manual' Australian Signals Directorate, November 2017, at https://acsc.gov.au/infosec/ism/index.htm

ISO (2005) 'Information Technology - Code of practice for information security management', International Standards Organisation, ISO/IEC 27002:2005

ISO (2008) 'Information Technology - Security Techniques - Information Security Risk Management' ISO/IEC 27005:2008

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

Knight W. (2017) 'The Dark Secret at the Heart of AI' 11 April 2017, MIT Technology Review https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Leenes R. & Lucivero F. (2014) 'Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design' Law, Innovation and Technology 6, 2 (2014) 193-220

Lessig L. (1999) 'Code and Other Laws of Cyberspace' Basic Books, 1999

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

McCarthy J., Minsky M.L., Rochester N. & Shannon C.E. (1955) 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence' Reprinted in AI Magazine 27, 4 (2006), at https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1904/1802

Manwaring K. & Clarke R. (2015) 'Surfing the third wave of computing: a framework for research into eObjects' Computer Law & Security Review 31,5 (October 2015) 586-603, PrePrint at http://www.rogerclarke.com/II/SSRN-id2613198.pdf

Martins L.E.G. & Gorschek T. (2016) 'Requirements engineering for safety-critical systems: A systematic literature review' Information and Software Technology Journal 75 (2016) 71-89

Maschmedt A. & Searle R. (2018) 'Driverless vehicle trial legislation ‚Äì state-by-state' King & Wood Malleson, February 2018, at https://www.kwm.com/en/au/knowledge/insights/driverless-vehicle-trial-legislation-nsw-vic-sa-20180227

Mayer-Schonberger V. & Cukier K. (2013) 'Big Data: A Revolution That Will Transform How We Live, Work and Think' John Murray, 2013

MS (2018) 'Microsoft AI principles' Microsoft, August 2018, at https://www.microsoft.com/en-us/ai/our-approach-to-ai

Newcomer E. (2018). 'What Google's AI Principles Left Out: We're in a golden age for hollow corporate statements sold as high-minded ethical treatises' Bloomberg, 8 June 2018, at https://www.bloomberg.com/news/articles/2018-06-08/what-google-s-ai-principles-left-out

NIST (2012) 'Guide for Conducting Risk Assessments' National Institute of Standards and Technology, Special Publication SP 800-30 Rev. 1, September 2012, at http://csrc.nist.gov/publications/nistpubs/800-30-rev1/sp800_30_r1.pdf

OTA (1977) 'Technology Assessment in Business and Government' Office of Technology Assessment, NTIS order #PB-273164', January 1977, at http://www.princeton.edu/~ota/disk3/1977/7711_n.html

Pagallo U. (2016). 'Even Angels Need the Rules: AI, Roboethics, and the Law' Proc. ECAI 2016

Palmerini E. et al. (2014). 'Guidelines on Regulating Robotics Delivery' EU Robolaw Project, September 2014, at http://www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922.pdf

Pichai S. (2018) 'AI at Google: our principles' Google Blog, 7 Jun 2018, at https://www.blog.google/technology/ai/ai-principles/

PoAI (2018) 'Our Work (Thematic Pillars)' Partnership on AI, April 2018, at https://www.partnershiponai.org/about/#pillar-1

Pouloudi A. & Whitley E.A. (1997) 'Stakeholder Identification in Inter-Organizational Systems: Gaining Insights for Drug Use Management Systems' European Journal of Information Systems 6, 1 (1997) 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

Rayome A.D. (2017) 'Guiding principles for ethical AI, from IBM CEO Ginni Rometty', TechRepublic (17 January 2017), at https://www.techrepublic.com/article/3-guiding-principles-for- ethical-ai-from-ibm-ceo-ginni-rometty/

Russell S.J. & Norvig P. (2009) 'Artificial Intelligence: A Modern Approach' Prentice Hall, 3rd edition, 2009

Schellekens M. (2015) 'Self-driving cars and the chilling effect of liability law' Computer Law & Security Review 31, 4 (Jul-Aug 2015) 506-517

Scherer M.U. (2016) 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' Harvard Journal of Law & Technology 29, 2 (Spring 2016) 353-400, at http://euro.ecom.cmu.edu/program/law/08-732/AI/Scherer.pdf

Selbst A.D. & Powles J. (2017) 'Meaningful information and the right to explanation' International Data Privacy Law 7, 4 (November 2017) 233-242, at https://academic.oup.com/idpl/article/7/4/233/4762325

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/

TvH (2006) 'Telstra Corporation Limited v Hornsby Shire Council' NSWLEC 133 (24 March 2006), esp. paras. 113-183, at http://www.austlii.edu.au/au/cases/nsw/NSWLEC/2006/133.htm

UGU (2017) 'Top 10 Principles for Ethical AI' UNI Global Union, December 2017, at http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

Vellinga N.E. (2017) 'From the testing to the deployment of self-driving cars: Legal challenges to policymakers on the road ahead' Computer Law & Security Review 33, 6 (Nov-Dec 2017) 847-863

Villani C. (2017) 'For a Meaningful Artificial Intelligence: Towards a French and European Strategy' Part 5 - What are the Ethics of AI?, Mission for the French Prime Minister , March 2018, pp.113-130, at https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

Villaronga E.F. & Roig A. (2017) 'European regulatory framework for person carrier robots' Computer Law & Security Review 33, 4 (Jul-Aug 2017) 502-220

Wachter S. & Mittelstadt B. (2019) 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' Forthcoming, Colum. Bus. L. Rev. (2019), at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3248829

Wachter S., Mittelstadt B. & Floridi L. (2017) 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' International Data Privacy Law 7, 2 (May 2017) 76-99, at https://academic.oup.com/idpl/article/7/2/76/3860948

Warwick K. (2014) 'The Cyborg Revolution' Nanoethics 8, 3 (Oct 2014) 263-273

WH (2018) 'Summary of the 2018 White House Summit on Artificial Intelligence for American Industry' Office of Science and Technology Policy, White House, May 2018, at https://www.whitehouse.gov/wp-content/uploads/2018/05/Summary-Report-of-White-House-AI-Summit.pdf

Wright D. & De Hert P. (eds) (2012) 'Privacy Impact Assessments' Springer, 2012

Wyndham J. (1932) 'The Lost Machine' (originally published in 1932), reprinted in A. Wells (Ed.) 'The Best of John Wyndham' Sphere Books, London, 1973, pp. 13- 36, and in Asimov I., Warrick P.S. & Greenberg M.H. (Eds.) 'Machines That Think' Holt, Rinehart, and Wilson, 1983, pp. 29-49

Zhaohui W. et al. (2016) 'Cyborg Intelligence: Recent Progress and Future Directions' IEEE Intelligent Systems 31, 6 (Nov-Dec 2016) 44-50

Acknowledgements

This paper has benefited from feedback from multiple colleagues, and particularly Peter Leonard of Data Synergies and Prof. Graham Greenleaf and Kayleen Manwaring of UNSW. I first applied the term 'intellectics' during a presentation to launch a Special Issue of the UNSW Law Journal in Sydney in November 2017.

Author Affiliations

Roger Clarke is Principal of Xamax Consultancy Pty Ltd, Canberra. He is also a Visiting Professor in Cyberspace Law & Policy at the University of N.S.W., and a Visiting Professor in the Research School of Computer Science at the Australian National University. He has also spent many years on the Board of the Australian Privacy Foundation, and is Company Secretary of the Internet Society of Australia.

Access

Personalia **Photographs** **Statistics**

Xamax Consultancy Pty Ltd
ACN: 002 360 456
78 Sidaway St, Chapman ACT
2611 AUSTRALIA
Tel: +61 2 6288 6916

Created: 11 July 2018 - Last Amended: 3 October 2018 by Roger Clarke - Site Last Verified: 15 February 2009
This document is at www.rogerclarke.com/EC/GAIF.html

Responsible AI Technologies, Artefacts, Systems and Applications

50 Principles

This document reproduces Appendix 1 of Clarke (2018a)

The following Principles are intended to be applied by the entities responsible for all phases of AI research, invention, innovation, dissemination and application. The cross-references are to the sources on 'Ethical Analysis and IT' sources (Clarke 2018b – E) and of 'Principles for AI' (Clarke 2018c – P).

1. Evaluate Positive and Negative Impacts

1.1 Conceive and design only after ensuring adequate understanding of purposes and contexts (E4.3, P5.3, P6.21, P7.1, P15.7)

1.2 Justify objectives (E3.25)

1.3 Demonstrate the achievability of postulated benefits
(Not found in any of the documents, but a logical pre-requisite)

1.4 Conduct impact assessment (E7.1, P3.12, P4.1, P4.2, P6.21, P11.8)

1.5 Publish sufficient information to stakeholders to enable them to conduct impact assessment (E7.3, P3.7, P4.1, P8.3, P8.4, P8.7)

1.6 Conduct consultation with stakeholders and enable their participation (E5.6, E7.2, P3.7, P8.6, P8.7)

1.7 Justify negative impacts on individuals ('proportionality') (E3.21, E7.4, E7.5)

1.8 Consider alternative, less harmful ways of achieving the same objectives (E3.22)

2. Complement Humans

2.1 Design as an aid, for augmentation, collaboration and inter-operability (P4.5, P5.1, P9.1, P9.8 .P14.2, P14.4)

2.2 Avoid design for replacement of people by independent devices, except in circumstances in which artefacts are demonstrably more capable than people, and even then ensuring that the result is complementary to human capabilities (P5.1)

3. Ensure Human Control

3.1 Ensure human control over AI-based artefacts and systems (E4.2, E6.1, E6.8, E6.19, P1.4, P2.1, P4.2, P5.2, P6.16, P8.4, P9.3, P12.1, P13.5, P15.4)

3.2 In particular, ensure control over autonomous behaviour of AI-based artefacts and systems (E8.1, P8.4, P10.2, P11.4)

3.3 Respect each person's autonomy, freedom of choice and self-determination (E2.1, E5.3, P3.3, P9.7, P11.3)

3.4 Ensure human review of inferences and decisions prior to acting on them (E3.11)

3.5 Respect people's expectations in relation to personal data protections (E5.6), incl.:
- awareness of data-usage (E3.6)
- consent (E3.7, E3.28, E5.3, P3.11, P4.6)
- data minimisation (E3.9)
- public visibility and consultation (E3.10, E7.2), and
- relationship of data-usage to the data's original purpose (E3.27)

3.6 Avoid deception of humans (E4.4, E6.20, P2.5)

3.7 Avoid services being conditional on the acceptance of AI-based artefacts and systems (P4.5)

4. Ensure Human Safety and Wellbeing

4.1 Ensure people's physical health and safety ('nonmaleficence')
(E2.2, E3.1, E4.1, E4.3, E5.4, E6.2, E6.9, E6.13, E6.14, E6.18, P1.2, P1.3, P2.1, P3.2, P3.6, P3.9, P3.12, P4.3, P4.9, P6.6, P8.4, P9.4, P10.2, P11.4, P13.5, P14.1, P15.3)

4.2 Ensure people's psychological safety (E3.1, E6.9, E6.13), by avoiding negative effects on any individual's mental health, inclusion in society, worth, standing in comparison with other people, or emotional state (E5.4, E6.3)

4.3 Ensure people's wellbeing ('beneficence')
(E2.3, E3.20, E5.5, P3.1, P3.4, P6.1, P6.14, P6.15, P8.6, P11.6, P12.2, P13.1, P15.1, P16.4)

4.4 Mitigate negative consequences (E3.24, E7.6, E6.21, E10.4)

4.5 Avoid violation of trust (E3.3)

4.6 Avoid the manipulation of vulnerable people (E4.4, P4.5, P4.9), including taking advantage of individuals' tendency to addiction, e.g. to gambling (E6.3)

5. Ensure Consistency with Human Values and Human Rights

5.1 Ensure compliance with human rights laws (E4.2, P3.5, P3.9, P4.3)

5.2 Be just / fair / impartial and treat individuals equally (E2.4, E3.16, E3.29, P3.4)

5.3 Avoid unfair discrimination and bias, not only where it is legally procribed but also where it is publicly unacceptable
(ICCPR Arts. 2.1, 3, 26 and 27, E3.16, P3.4, P4.5, P11.5, P15.2, P16.1)

5.4 Avoid restrictions on freedom of movement (ICCPR 12, P6.13)

5.5 Avoid interference with privacy, family, home or reputation
(ICCPR 17, E5.6, P3.11, P6.12, P8.4, P9.6, P13.3, P15.5)

5.6 Avoid interference with the rights of freedom of information, opinion and expression (ICCPR 19, P4.6)

5.7 Avoid interference with the right of freedom of assembly (ICCPR 21, P6.13)

5.8 Avoid interference with the right of freedom of association (ICCPR 22, P6.13)

5.9 Avoid interference with the rights to participation in public affairs and access to public service (ICCPR 25, P6.13)

6. Embed Quality Assurance

6.1 Invest in quality assurance (E6.2, P4.2, P15.6)

6.2 Ensure effective, efficient and adaptive performance of intended functions
(E6.11, P1.6)

6.3 Ensure security safeguards against inappropriate modification to and deletion of sensitive data (E3.15)

6.4 Ensure justification of the use of sensitive data (E3.26, E7.4)

6.5 Ensure data quality and data relevance (P10.3, P11.2)

6.6 Deal fairly with people (faithfulness, fidelity) (E2.5, E3.2)

6.7 Avoid invalid and unvalidated techniques (E3.5, P7.7)

6.8 Test for result validity (E3.5, P7.7. P9.2)

6.9 Impose controls in order to ensure that safeguards are operative and effective
(E7.7, P10.2)

6.10 Conduct audits of safeguards and controls (E7.8, P9.3)

7. Deliver Transparency and Auditability

7.1 Ensure that the fact that the process is AI-based is transparent to all stakeholders (E4.4, P4.8)

7.2 Ensure that the means whereby inferences are drawn, decisions made and actions are taken are logged and can be reconstructed (E6.6, P2.4, P4.8, P5.2, P6.7, P7.4, P7.6, P8.2. P9.2, P11.1, P11.2, P13.2, P16.2, P16.3)

7.3 Ensure people are aware of inferences and how they were reached (E3.12, P2.4)

8. Exhibit Robustness and Resilience

8.1 Deliver and sustain appropriate security safeguards against compromise of intended functions arising from both passive threats and active attacks (E4.3, E6.11, P1.4, P1.5, P4.9, P6.6, P8.4, P9.5)

8.2 Deliver and sustain appropriate security safeguards against inappropriate access to sensitive data arising from both passive threats and active attacks (E3.15, E6.5, E6.10, P3.11, P9.6)

8.3 Conduct audits of justification, proportionality, transparency, mitigation measures and controls (E7.8, E8.4)

8.4 Ensure resilience, in the sense of prompt and effective recovery from incidents

9. Ensure Accountability for Legal and Moral Obligations

9.1 Ensure that the responsible entity is apparent or can be readily discovered by any party (E4.5, E6.4, P2.3, P3.8, P4.7, P8.5, P12.3)

9.2 Ensure that effective remedies exist, in the form of complaints processes, appeals processes, and redress where harmful errors have occurred (ICCPR 2.3, E3.13, E3.14, E7.7, P3.11, P4.7, P7.2, P8.7, P9.9, P10.5, P11.9, P16.3)

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

10.1 Ensure that complaints, appeals and redress processes operate effectively (ICCPR 2.3, E7.7)

10.2 Comply with external complaints, appeals and redress processes and outcomes (ICCPR 14), including, in particular, provision of timely, accurate and complete information relevant to cases

---

References

Clarke R. (2018a) 'Guidelines for the Responsible Business Use of AI – Foundational Working Paper' Xamax Consultancy Pty Ltd, October 2018, at http://www.rogerclarke.com/EC/GAIF.html

Clarke R. (2018b) 'Ethical Analysis and Information Technology' Xamax Consultancy Pty Ltd, October 2018, at http://www.rogerclarke.com/EC/GAIE.html

Clarke R. (2018c) 'Principles for AI: A SourceBook' Xamax Consultancy Pty Ltd, October 2018, at http://www.rogerclarke.com/EC/GAIP.html

---