COMPUTER LAW & SECURITY REVIEW 35 (2019) 423–433

Available online at www.sciencedirect.com

ScienceDirect

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

Computer Law & Security Review

Why the world wants controls over Artificial Intelligence

Roger Clarke[a,b,c,*]

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

A B S T R A C T

This article reviews the nature, the current state and possible future of Artificial Intelligence (AI). AI is described both in the abstract and in four forms that are currently evident not only in laboratories but also in real-world applications. Clarity about the public's concerns is sought by articulating the threats that are inherent within AI. It is proposed that AI needs to be re-conceived as 'complementary artefact 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'. This article lays a foundation for two further articles on how organisations can adopt a responsible approach to AI, and how an appropriate regulatory scheme for AI can be structured.

1. Introduction

Since the conception of Artificial Intelligence (AI) in the early post-World War II period, there have been sporadic surges in marketing fervour for various flavours of it. Its aura of mystery and confusion, compounded by a considerable amount of over-claiming, has stimulated periods of public enthusiasm interspersed with 'winters of discontent'.

Several forms of AI are currently being vigorously promoted, and are attracting attention from investors, user organisations, the media and the public. However, along with their promises, they bring major challenges in relation to understandability, control and auditability.

To date, public understanding of AI has been marketer-driven and superficial. This is a perfect breeding-ground for mood-swings, between euphoric and luddite. Many people are wary about AI inherently undermining accountability and stimulating the abandonment of rationality. Cautionary voices have included cosmologist Stephen Hawking (Cellan-Jones 2014), Microsoft billionaire Bill Gates (Mack 2015), and technology entrepreneur Elon Musk (Sulleyman 2017).

Meanwhile, less prominent people are suffering from unreasonable inferences, decisions and actions by AI-based artefacts and systems. One form of harm is unfair and effectively unappealable decisions by government agencies about benefits and penalties, by financiers about credit-granting, by insurers, and by employers. In addition, instances are accumulating of physical harm arising from autonomous acts by artefacts such as cars and aircraft. Aggrieved victims are likely to strike back against the technologies and their purveyors.

This article is addressed to a readership that is technically literate, socially aware, and concerned with technology policy and law. It accordingly assumes moderate familiarity with the topic. It commences with brief overviews of AI and of several key forms of it. The aim is to enable delineation of the threats that accompany AI's promises, and that give rise to the need

* Corresponding authors at: Xamax Consultancy Pty Ltd, 78 Sidaway St, Chapman ACT 2611, Australia.
E-mail address: Roger.Clarke@xamax.com.au

https://doi.org/10.1016/j.clsr.2019.04.006
0267-3649/

for responsibility to be shown in relation to its development and deployment.

2. Artificial Intelligence

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 and Norvig 2009, pp. 16–28, Boden 2016, pp.1–20) identify multiple strands, but also multiple re-visits to much the same territory, and a considerable degree of creative chaos.

Many attempts have been made to distill out the sense in which the juxtaposition of the two words is to be understood. Conventionally (Albus 1991, Russell & Norvig 2003, McCarthy 2007):

***Intelligence** is exhibited by an artefact if it*

(1) *Evidences perception and cognition of relevant aspects of its environment*
(2) *Has goals; and*
(3) *Formulates actions towards the achievement of those goals*

The word 'artificial' implies 'artefactual' or 'human-made'. Its conjunction with 'intelligence' has imbued it with competing ideas about whether the yardstick is 'equivalent to human', 'different from human' or 'superior to human'.

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 even crossed the border into the realms of alchemy. The exaggerations have resulted in under-delivery and a cyclical 'boom and bust' pattern, with research funding being sometimes very easy to obtain, and sometimes very difficult, depending on whether the focus at the time was on the hyperbole or on the very low delivery-rate against promises.

Part of AI's image-problem is that successes deriving from what began as AI research have shed the name. In a quotation widely-attributed to John McCarthy, "As soon as it works, no-one calls it AI anymore". 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, optical character recognition (OCR), automated number-plate recognition (ANPR), and object and facial recognition. Game-playing, particularly of chess and go, has surpassed human-expert levels, and provided entertainment value and spin-offs. It is as yet unclear, however, whether AI-based game-playing has provided the breakthroughs towards posthumanism that its proponents appeared to be claiming for it.

Organisations, in business and government alike, need to identify AI technologies that have practical value, and devise ways to apply them so as to 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 piloted and empirically tested in real-world contexts, under sufficiently controlled conditions that the risks have been able to be identified, assessed and then managed.

The scope addressed in this article is very broad, in terms of both technologies and applications, but it excludes design and use for warfare or armed conflict. It is, however, intended to include applications to civil law enforcement and domestic national security, i.e. safeguards for the public, for infrastructure, and for public figures. The following section undertakes brief scans of several current technologies that are within the field of view.

3. AI exemplars

AI's scope is broad, and contested. This section identifies several technologies that have current relevance. That relevance derives in part from claims of achievement of progress and benefits, and in part from media coverage resulting in awareness among organisations' executives and staff and the general public. In addition to achieving a level of adoption, each faces degrees of technical challenge, public scepticism and resistance.

The following sub-sections briefly review four AI technologies, with a view to enabling commonalities to emerge among the diversity of features.

3.1. *Robotics*

The two foundational elements of robotics are programmability, implying computational or symbol-manipulative capabilities that a designer can combine as desired (a robot is a com-

puter); and mechanical capability, whereby inbuilt actuators influence its environment (a robot is a machine). A comprehensive design also requires sensors to acquire data from the robot's environment (Arkin 1998).

Robotics has built on its earlier successes in controlled environments such as the factory floor and the warehouse, and is now in direct contact with the public. Some applications are non-obvious, such as low-level control over the attitude, position and course of craft both on or in water and in the air. Others are more apparent. The last few years have seen a great deal of activity in relation to self-driving vehicles (Paden et al. 2016), variously on rails and otherwise, in controlled environments such as mines, quarries and dedicated tram, train and 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. The appearance of robotic technologies in public spaces has attracted attention and rejuvenated concerns about their impacts and implications.

3.2. *Cyborgisation*

The term 'cyborg' was coined from 'cybernetic organism' to refer to a technologically enhanced human being, originally in the context of survival in extraterrestrial environments (Clynes and Kline 1960). 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 (Mann and Niedzviecki 2001; 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 research in the field is emergent (Zhaohui et al. 2016). Substantial progress with medical implants (Bhunia et al. 2015) suggests that these technologies have the prospect of becoming a flash-point for public concerns, because they involve direct intervention with the human body.

3.3. *Rule-Based expert systems*

Computing applications for drawing inferences from data began with hard-wired, machine-level and assembler languages (1940–1960), but made great progress with higher-level, imperative languages (indicatively, 1960–1990), particularly those that enabled the coding of genuinely 'algorithmic' programs, such as ForTran (Formula Translator). This 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, which embody no explicit 'problem' or 'solution'. Rule-based expert systems involve the representation of human expertise as statements about relationships between 'antecedent' and 'consequent' variables, in the form 'if-then'. The relationships may be theoretically-based and/or empirically-derived, mere heuristics / 'rules of thumb', or just hunches. When software that embodies sets of rules is provided with data, it applies the rules to that data, and draws inferences (Giarratano and Riley 1998). Frequently-cited applications include decisions about an individual's eligibility for citizenship or credit-worthiness and about the legality or otherwise of an act or practice.

Unlike algorithmic or procedural approaches, rule-based expert systems embody no conception of either a problem or a solution. A rule-base merely describes a problem-domain in a form that enables inferences to be drawn from it (Clarke 1991). In order to understand the rationale underlying an inference, a human needs access to the rules that were 'fired', and the data that gave rise to their invocation. This may or may not be supported by the software. Even if access is supported, this may or may enable human understanding of the rationale underlying the inference, and whether or not the inference is reasonable in the circumstances.

3.4. *AI/ML/Neural networks*

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 is 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, in particular the weight associated with each arc, 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 the initial set of weights on the connections between nodes.

Unlike previous techniques for developing software, neural networking approaches need 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 what the analyst has determined to be the key elements and processes of a real-world system, a neural network model may be merely lists of input variables and output variables (and, in the case of 'deep' networks, one or more levels of intermediary variables). To the extent that a model exists, in the sense of a representation of the real world, it is implicit rather than express. The weights 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.

Enthusiasts see great prospects in neural network techniques, e.g. "There has been a number of stunning new results with deep-learning methods ... The kind of jump we are seeing in the accuracy of these systems is very rare indeed" (Markoff 2012). They claim that noisy and error-ridden data presents no problems, provided that there's enough of it. They also claim that the techniques have a wide range of application areas. Sceptics, on the other hand, perceive that the techniques' proponents overlook serious weaknesses (Marcus 2018), and in effect treat empiricism as entirely dominating theory. Combining these issues with questions about the selectivity, accuracy and compatibility of the data gives rise to considerable uncertainty about the techniques' degree of affinity with the real world circumstances to which they are applied.

Inferences drawn using neural networking inevitably reflect errors and biasses inherent in the implicit model, in the selection of real-world phenomena for which data was created, in the selection of the training-set, and in the particular learning algorithms used to develop the application. 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 of the compatibility among them all, 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 need to be regarded as unproven techniques that harbour considerable dangers to the interests of organisations and their stakeholders.

3.5. *Commonalities among these AI Exemplars*

The four AI technologies outlined here exhibit considerable differences, but also some commonalities. One important common factor is the lack of transparency about the means whereby inferences are drawn, decisions are made, and (in two cases) actions are taken. The fog may be so dense that no scope exists for human understanding of the process, and there may even be no rationale and no means of reconstructing one. Another common feature is intrusiveness into human affairs, in some cases by the very nature of the technology, and in others as a result of the contexts within which they are applied. Proponents of the technologies also make assumptions about the nature of the data on which they depend, often without checking that the assumptions are justified, and without meaningful consideration of the implications if they turn out to be wrong.

4. The threats inherent in AI

The characteristics of AI, and of the four mainstream forms outlined in the previous section, give rise to a wide array of serious concerns about AI's impacts and implications (e.g. Scherer 2016, esp. pp. 362–373, Yampolskiy and Spellchecker 2016, Duursma 2018). Many of the concerns may be keenly-felt, but are vague, such as the disruption of work-based income-distribution, the imposition of predestination on individuals, the dominance of collectivism over individualism, the undermining of human rights, the disruption of culture, the dominance of the powerful over the weak, and the risk of undermining the meaningfulness of human life.

The following is proposed as an expression of concern that has the capacity to provide guidance for responsible behaviour:

> *AI gives rise to errors of inference, of decision and of action, which arise from the more or less independent operation of artefacts, for which no rational explanations are available, and which may be incapable of investigation, correction and reparation*

Even this expression requires unpacking, however, in order to identify problems that can be addressed by the crafting of safeguards. The following sections discuss five factors that underlie the above expression of the concerns about AI. The first consideration is the extent of human delegation to artefacts. This is followed by a consideration of assumptions about data and about the processes used to draw inferences from data, and of the opaqueness of those inferences. The final factor examined is the failure to sheet home responsibilities to the entities involved in the AI industry supply chain.

4.1. *Artefact 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.

A common feature of the four AI technologies discussed earlier is that, to a much greater extent than in the past, software is drawing inferences, making decisions, and taking action. Put another way, artefacts are being imbued with a much greater degree of autonomy than was the case in the past.

Artefact autonomy may merely comprise a substantial repertoire of pre-programmed stimulus-response relationships. Alternatively, it may extend to the capacity for auto-adaptation of aspects of those relationships, or for the creation of new relationships. For example, where machine-learning is applied, the stimulus-response relationships

change over time depending on the cases handled in the intervening period.

As a result of emergent artefact autonomy, humanity is in the process of delegating not to humans, but to human inventions. This gives rise to uncertainties whose nature is distinctly different from prior and well-trodden paths of human and organisational practice. A further relevant factor is that autonomous artefacts have a high likelihood of stimulating repugnance among a proportion of the public, and hence giving rise to luddite 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 process 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 gravity-relative 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 sustain progress towards a human-pre-determined or human-amended destination, or 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, and some drones that lose contact with their pilot can decide when and where to land. To date, high-order activities that are seldom if ever autonomous include decisions about the mission objective and when to take off, and adjustments to the objective and destination.

Artefact autonomy can be absolute, but is more commonly qualified, in that a human – or perhaps some superordinate artefact – can exercise some degree of control over the artefact's behaviour. Table 1 draws on and simplifies various models that provide structure to that relationship, including Armstrong (2010, p.14), Clarke (2014a, Table 1) and Sheridan & Verplank (1978, Table 8.2, pp. 8-17-8.19) as interpreted by Robertson et al. (2019, Table 1).

It is readily argued that the degree of autonomy granted to artefacts needs to reflect the layer at which the particular function is operating. The sequence in which the alternatives are presented in Table 1 corresponds with those layers. At the lowest level (7), the rapidity with which analysis, decision and action need to be undertaken may preclude conscious human involvement. At the other extreme (1), artefacts lack the capability to deal with the complexities, ambiguities, variability, fluidity, value-content and value-conflicts inherent in important real-world decision-making (Dreyfus 1992). There are circumstances (5–6) in which it is appropriate to enable autonomous behaviour by artefacts subject to human interruption or override. In other circumstances (2–4), the appropriate approach is for the artefact to provide decision support to humans, through analysis, advice and/or recommendation.

There appears to be *de facto* public acceptance of the notion of delegation of low-level, real-time functions to artefacts. Even at that level, however, AI is adding a further level of mystery. It remains to be seen whether the public will continue to accept inexplicable events resulting in aircraft and driverless-vehicle incidents. Following the crash of a second Boeing 737 Max in early 2019, the US President voiced a popular sentiment, to the effect that pilots should be professionals who can easily and quickly take control of their aircraft. That portends an edict that robot autonomy, at least for passenger aircraft, will be limited to revocable autonomy (5–6), with layer 7 prohibited. In respect of less structured decisions, there seems little prospect of public acceptance even of revocable automated decision-making.

IEEE, even though 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 'artificial', and prefers the term 'Autonomous and Intelligent Systems (A/IS)'.

4.2. *Inappropriate assumptions about data*

An artefact's awareness of its environment depends on data variously provided to it and acquired by its sensors. Any deficiencies in the quality of that data undermine the appropriateness of the artefact's inferences, decisions and actions.

Data quality is a function of a large set of factors (Wang and Strong 1996; Clarke 2016b). Beyond validity, accuracy, precision, timeliness, completeness, and general and specific relevance, the correspondence of the data with the real-world phenomena that the process assumes it to represent depends

Table 1 – Degrees of autonomy.

		Function of the artefact	Function of the controller
Decision support system	1.	NIL	**Act**
	2.	Analyse options	**Decide among options**
	3.	Advise re options	**Decide among options**
	4.	Recommend action	**Approve/Reject recommended action**
Decision system	5.	**Notify an Impending action**	Override/Veto an impending action
	6.	**Act and inform**	Interrupt/Suspend/Cancel an action
	7.	**Act**	NIL

on appropriate identity association, attribute association and attribute signification.

Where data is drawn from multiple sources, definitional and quality consistency among those sources is almost inevitably a limiting factor, yet it is seldom considered (Widom 1995). Data scrubbing (or 'cleansing') may be applied; but this is a dark art, and most techniques generate some errors in the process of correcting others (Mueller and Freytag 2003). Further, attention has already been drawn to the often-expressed claim that, with sufficiently large volumes of data, the impacts of low data, matching and scrubbing quality automatically smooth themselves out. This is a justifiable claim in specific circumstances, but in most cases is a magical incantation that does not hold up under cross-examination (boyd and Crawford 2012).

4.3. *Inappropriate assumptions about the inferencing process*

Endeavours are being made to apply robotics outside the controlled environments in which they have enjoyed success (factories, warehouses, and thinly human-populated mining sites) to contexts in which there is much more variability and unpredictability, and much less structure (such as public roads, households, and human care applications).

In the case of the flying robots popularly called drones, considerable challenges confront the design and deployment even of a generally applicable process for safe landing when communications are lost with the pilot, let alone collision-detection capabilities, far less collision-avoidance functionality. Yet these are processes that are expectations and even legal obligations in current, human-operated activities, and hence pre-conditions for AI-based substitutes.

Where AI technologies depend on the drawing of inferences from data, confidence is needed that, in each case, and before reliance is placed upon it, the inferencing process's applicability to the particular problem-category or problem-domain has been demonstrated – preferably both theoretically and empirically.

A further issue is the suitability of the available data as input to the particular inferencing process. A great deal of data is on nominal scales (which merely distinguishes categories). Some is on ordinal scales (implying a structured relationship between categories, such as 'good, better, best'), and some is on cardinal scales (with equal intervals between the categories, such as temperature expressed in degrees Celsius). Only a limited range of analytical tools is available for data on such scales. Most of the powerful statistical tools applied by data analysts assume that all of the data is on ratio scales (which feature equal intervals and a natural zero, such as degrees Kelvin). Many analyses abuse the rules of statistics by applying techniques inappropriately. Mixed-mode data (i.e. where the various items of data are on different kinds of scale) is particularly challenging to deal with. Further, most tools cannot cope with missing values, and hence more or less arbitrary values need to be invented. Given the problems that need to be overcome, it is highly inadvisable for inferencing mechanisms to be relied upon as a basis for decision-making that has material consequences, unless and until their applicability to the data in question has been subjected to independent analysis and certification.

Of particular concern are assertions that empirical correlation unguided by theory is enough, and that rational explanation is a luxury that the world needs to learn to live without. These cavalier claims are published not only by excitable journalists but also by influential academics (Anderson 2008; LaValle et al. 2011; Mayer-Schoenbeger and Cukier 2013).

4.4. *Opaqueness of the inferencing process*

Some forms of AI, importantly including neural networking, are largely empirical rather than based on an established theory. Moreover, where they embody any form of machine learning, their performance may vary over time even though the context appears unchanged. Some other AI technologies are built on a stronger theoretical base, but are complex and multi-layered. These characteristics make it difficult for humans to grasp how AI does what it does, and to explain and understand the inferences it draws, the decisions it makes, and the actions it takes (Burrell 2016; Knight 2017).

This lack of transparency gives rise to many further features, summarised in Table 2. Not all of these may be evident in any given situation, but all of them may have serious consequences for individuals and organisations.

Where decision transparency is absent, the accountability of organisations for their decisions is undermined. Where entities are secure in the knowledge that blame cannot be sheeted home to them, irresponsible behaviour is inevitable. Under threat are the established principles of evaluation, fairness, proportionality, evidence-based decision-making, and the capacity to challenge decisions (APF 2013).

There is increasing public pressure for explanations to be provided for decisions that are adverse to the interests of individuals and of small business. The responsibility of decision-makers to provide explanations has always been implied by the principles of natural justice and procedural fairness. In many jurisdictions, administrative law imposes specific requirements on government agencies. In the private sector as well, organisations are gradually becoming subject to legal provisions. In the EU, since mid-2018, access must be provided to "meaningful information about the logic involved", "at least in" the case of automated decisions (GDPR 2018, Articles 13.2(f), 14.2(g) and 15.1(h), Selbst and Powles 2017). The scope and effectiveness of these provisions is as yet unclear. One interpretation is that "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 … [Hence] a new data protection right, the 'right to reasonable inferences', is needed" (Wachter and Mittelstadt 2019).

4.5. *Irresponsibility*

A further factor is at work in undermining accountability. There has to date been inadequate discrimination among the various stages of the supply-chain from laboratory experiment to deployment in the field. This leads to a failure to assign responsibilities to the various categories of entities.

Table 2 – Implications of the lack of process transparency.

- ***A-rationality***
  A description of how and why an outcome came about may not exist, and an *ex post facto* rationalisation of it may not be able to be constructed, with the result that no humanly-understandable explanation can be provided
- ***Unreplicability***
  A process performed by means of AI may not be able to be repeated, which undermines the scope for investigation and reconstruction of the sequence of events
- ***Unauditability***
  A process may not be able to be checked by an independent party such as an auditor, judge or coroner, because records of the initial state, intermediate states and triggers for transitions between states, may not exist and may not be able to be re-constructed
- ***Uncorrectability***
  Even where an outcome appears to be in error, the factors that gave rise to it may not be discoverable, and, in the absence of an explanation, undesired actions may not be correctable
- ***Unaccountability***
  Even if an entity has nominal responsibility for a decision or action, it may escape accountability, perhaps on grounds similar to *force majeure*, i.e. AI's opaqueness may be seen as a force that is beyond the capacity of a human entity or organisation to cope with, thereby absolving it of responsibility

In Table 3, the AI supply-chain is depicted as a succession of phases, from laboratory experiment to deployment in the field. Distinctions are drawn among technology, artefacts that embody the technology, systems that incorporate the artefacts, and applications of those systems. Appropriate responsibilities can then be assigned to, successively, researchers, inventors, innovators, purveyors, and users. Each of these categories of entity bears moral responsibility for disbenefits arising from AI. Further, each of these categories of entity needs to be subject to legal constraints and obligations, commensurate with the roles that they play.

This section has sought to unbundle the many aspects of AI that embody threats, and that are at the heart of the public's demands for controls over AI. The following two articles in the series examine how organisations can exercise responsibility in the consideration of AI, and how a regulatory regime can be structured to ensure effective safeguards. The final section in this paper suggests that reconception of the field can be instrumental in assisting in the achievement of responsibility in relation to technology, artefacts and systems.

5. 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. After 65 years of confusion, it is high time that the key ideas were disentangled, and an interpretation adopted that can assist user organisations to appreciate the nature of the technologies, and then analyse those technologies' potential contributions and downsides.

Table 3 – 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	Users

This section suggests two conceptualisation that are intended to assist in understanding and addressing the technical, acceptance and adoption challenges.

5.1. *Complementary artefact intelligence*

If the intelligence that AI delivers is intended to be 'equivalent to human', some doubt has to be expressed about the value of the exercise. 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. If, on the other hand, the intelligence sought is in some way 'superior-to-human', the question arises 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 Artificial Intelligence is superior to human intelligence, can human intelligence reliably define what 'superior-to-human' intelligence 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, by working

with people and other artefacts. The following operational definition is proposed:

Complementary Artefact Intelligence:

(1) *does things well that humans do poorly or cannot do at all*
(2) *performs functions within systems that include both humans and artefacts; and*
(3) *interfaces effectively, efficiently and adaptably with both humans and artefacts*

A concept related to, but different from, 'complementary intelligence' is 'augmented intelligence' (Engelbart 1962; Mann 2001, but currently enjoying a revival). A fuller description of the concept that this section is addressing is as follows:

> *Complementary Artefact Intelligence refers to forms of Artefact Intelligence that are complementary to Human Intelligence, and that work with Human Intelligence synergistically, thereby producing a blend of human and artefact intelligence to which the term Augmented Intelligence is applied*

An alternative, imprecise but cute depiction is:

> *Human Intelligence*
> *+ Complementary Artefact Intelligence*
> *= Augmented Intelligence*

An important category of Complementary Artefact 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 exemplified by the distinction drawn in Table 1 above 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. MIT Media Lab's Joichi Ito has used the term 'extended intelligence' in a manner that links the notions of complementary artefact intelligence, augmented intelligence and responsible AI (Simonite 2018).

There are circumstances in which computer-based systems have clear advantages over humans, e.g. where significant computation is involved, and reliability, accuracy, and speed of inferencing, decision-making and/or action-taking are important. A pre-condition is, however, that a satisfactory structured process must exist. An alternative pre-condition may emerge, but is contentious. Some purely empirical techniques, and perhaps even heuristics ('rules of thumb'), may achieve widespread acceptance, e.g. if they are well-demonstrated to be more effective than either theory-driven approaches or human-performed decision-making.

Computer-based systems may have further advantages in relation to cost, and in relation to what in military contexts are referred to as "dull, dirty, or dangerous missions". Even where such superiority can be demonstrated, however, the need exists to shift discussion away from 'AI' to complementary intelligence, to technologies that augment human capabilities, and to systems that feature collaboration between humans and artefacts.

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

5.2. *Intellectics*

Robotics began with machines (in the sense of mechanical apparatus) 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 instead 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 'intellectics' has been previously used in a related but somewhat different 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 as follows:

> ***Intellectics** refers to a context in which artefacts go beyond merely drawing inferences from data, in that they take autonomous action in the real world*

In Table 1, decision systems were contrasted with decision support systems on the basis of the artefact's degree of autonomy. Table 4 identifies the forms that intellectics may take.

Table 4 – Forms of intellectics.

Full artefact autonomy
An artefact makes a decision, and takes action in the real world to give effect to that decision, without an opportunity for a human to prevent the action being taken
Revocable artefact autonomy
An artefact makes a decision, and informs a human controller that the action has been taken, and the human has the opportunity and capacity to interrupt the action
Overridable artefact autonomy
An artefact makes a decision, and informs a human controller that the action will be taken unless the human exercises their power to veto it, and the human has the opportunity and capacity to prevent the action

> ***The Threshold Test for Intellectics*** *is that, if the artefact cannot proceed with an action it has recommended unless the human exercises their power to accept the recommendation, then the artefact does not have autonomy*

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. Applying the notion of Intellectics has the benefit of bringing into much stronger focus the importance of assuring legitimacy of the data, of the inferencing technique, and of the inferences, decisions and actions.

In the case of inferencing based on neural networks, for example, major challenges that have to be satisfactorily addressed include the choice of learning algorithm, the availability and choice of sufficient training data, the quality of the training data, the significance of and the approaches adopted to data scrubbing and to empty cells within the training data, and the quality of the data to which the neural network is then applied (Clarke 2016a, 2016b).

6. Conclusions

This article has outlined AI, both in the abstract and through four exemplar technologies. That has enabled clarification of the threats inherent in AI, thereby articulating the vague but intense public concerns about the phenomenon.

This article has also proposed that the unserviceable notion of AI should be replaced by the notion of 'complementary artefact intelligence', and that the notion of robotics ('machines that think') is now much less useful than that of 'intellectics' ('computers that do'). In the near future, it may be possible to continue discussions using those terms. Currently, however, the mainstream discussion is about 'AI', and the further two articles in this series reflect that norm.

Sensor-computer-actuator packages are now generating a strong impulse for action to be taken in and on the real world, at the very least communicating a recommendation to a human, but sometimes generating a default-decision that is subject to being overridden or interrupted by a human, and even acting autonomously based on the inferences that software has drawn.

A power-shift towards artefacts is of enormous significance for humankind. It is also, however, a power-shift away from individuals and towards the mostly large and already-powerful organisations that control AI-based artefacts. Substantial pushback from the public needs to be anticipated. Existing regulatory arrangements need to be reviewed in light of the risks arising from AI. If adequate safeguards do not exist, new regulatory obligations will need to be imposed on organisations.

This article has identified a wide range of reasons why responsible behaviour by organisations in relation to AI is vital to the future for individuals, society and even humankind as a whole. The next article in the series examines how organisations can adapt their business processes, and apply a body of principles, in order to act responsibly in relation to AI technologies and AI-based artefacts and systems. The third article then addresses the question of how a regulatory regime can be structured, in order to encourage, and enforce, appropriate behaviour by all organisations.

Acknowledgements

This paper has benefited from feedback from multiple colleagues, and particularly Peter Leonard of Data Synergies, Prof. Graham Greenleaf and Kayleen Manwaring of UNSW and Prof Tom Gedeon of ANU. The comments of an anonymous referee were also helpful in ensuring clarification of key elements of the argument. I first applied the term 'intellectics' during a presentation to launch a Special Issue of the UNSW Law Journal in Sydney in November 2017.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.clsr.2019.04.006.

REFERENCES

Albus JS. Outline for a theory of intelligence. IEEE Trans Syst, Man Cybern 1991;21(3):473–509, at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.410.9719&rep=rep1&type=pdf.

Anderson C. The end of theory: the data deluge makes the scientific method obsolete. Wired Mag 2008;16(7), at http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

APF. Meta-principles for privacy protection. Australian Privacy Foundation, 2013, at https://privacy.org.au/policies/meta-principles/.

Arkin RC. Behavior-based robotics. MIT Press; 1998.

Armstrong AJ. Development of a Methodology for Deriving Safety Metrics for UAV Operational Safety Performance Measurement. Report, Master of Science in Safety Critical Systems Engineering, Department of Computer Science, York University, 2010, at http://www-users.cs.york.ac.uk/~mark/projects/aja506_project.pdf.

Bhunia S, Majerus SJA, Sawan M, editors. Implantable biomedical microsystems: design principles and applications. ScienceDirect 2015.

Bibel W. 'Intellektik' statt 'KI' — Ein ernstgemeinter Vorschlag. Rundbrief der Fachgruppe Kuenstliche Intelligenz in der Gesellschaft fuer Informatik; 1980;22.

Bibel W. The technological change of reality: opportunities and dangers. AI Soc 1989;3(2):117–32.

Boden M. AI: its nature and future. Oxford University Press; 2016.

boyd D, Crawford K. Critical questions for big data. Inf Commun Soc 2012;15(5):662–79, at http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878#.U_0X7kaLA4M.

Burrell J. How the machine 'thinks': understanding opacity in machine learning algorithms'. Big Data Soc 2016;3(1):1–12.

Cellan-Jones R. 'Stephen Hawking warns artificial intelligence could end mankind'. BBC News; 2014, at http://www.bbc.com/news/technology-30290540.

Clarke R. Knowledge-based expert systems: risk factors and potentially profitable application area. Xamax Consultancy Pty Ltd; 1989, at http://www.rogerclarke.com/SOS/KBTE.html.

Clarke R. A contingency approach to the application software generations. Database 1991;22(3):23–34, PrePrint at http://www.rogerclarke.com/SOS/SwareGenns.html.

Clarke R. Asimov's laws of robotics: implications for information technology in two parts. IEEE Comput 26, 12 (December 1993) 53–61 and 27,1 (January 1994) 57–66, at http://www.rogerclarke.com/SOS/Asimov.html.
Clarke R. 'Human-artefact hybridisation: forms and consequences'. Proceedings of the ARS Electronica 2005 symposium on hybrid - living in paradox, Linz, Austria; 2005, at http://www.rogerclarke.com/SOS/HAH0505.html.
Clarke R. Cyborg rights. IEEE Technol Soc 2011;30(3):49–57, at http://www.rogerclarke.com/SOS/CyRts-1102.html.
Clarke R. Understanding the drone epidemic. Comput Law Secur Rev 2014a;30(3):230–46, PrePrint at http://www.rogerclarke.com/SOS/Drones-E.html.
Clarke R. What drones inherit from their ancestors. Comput Law Secur Rev 2014b;30(3):247–62, PrePrint at http://www.rogerclarke.com/SOS/Drones-I.html.
Clarke R. Big data, big risks. Inf Syst J 2016a;26(1):77–90, PrePrint at http://www.rogerclarke.com/EC/BDBR.html.
Clarke R. Quality assurance for security applications of big data. Proceedings of the EISIC'16, Uppsala; 2016b. p. 17–19, PrePrint at http://www.rogerclarke.com/EC/BDQAS.html.
Clynes ME, Kline NS. Cyborgs and Space. Astronautics; 1960, pp. 26–27 and 74–75; reprinted in Gray, Mentor, and Figueroa-Sarriera, editors. 'The Cyborg Handbook' New York: Routledge, 1995, pp. 29-34.
Dreyfus HL. What computers still can't do: a critique of artificial reason. MIT Press; 1992.
Duursma. The risks of artificial intelligence. Studio OverMorgen; 2018, at https://www.jarnoduursma.nl/the-risks-of-artificial-intelligence/.
Engelbart DC. Augmenting human intellect: a conceptual framework. SRI Summary Report AFOSR-3223; 1962, at http://dougengelbart.org/content/view/138.
GDPR. General data protection regulation. Proceedings of the 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, 2018. http://www.privacy-regulation.eu/en/index.htm.
Giarratano JC, Riley G. Expert systems. 3rd Ed. PWS Publishing Co Boston; 1998.
IEEE. Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems (A/IS). IEEE; 2017, Version 2, at http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.
Knight W. The dark secret at the heart of AI. MIT Technology Review; 2017, at https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.
Kurzweil R. The singularity is near. Viking Books; 2005.
LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N. Big data, analytics and the path from insights to value. Sloan Manag Rev (Winter 2011 Res Feature) 2011. 21 December 2010, at http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/.
McCarthy J. What is artificial intelligence? Department of Computer Science, Stanford University; 2007, at http://www-formal.stanford.edu/jmc/whatisai/node1.html.
McCarthy J, Minsky ML, Rochester N, Shannon CE. (1955) A proposal for the dartmouth summer research project on artificial intelligence. Reprinted in AI Magazine 2006;27(4), at https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1904/1802.
Mack E. Bill Gates says you should worry about artificial intelligence. Forbes Magazine; 2015, at https://www.forbes.com/sites/ericmack/2015/01/28/bill-gates-also-worries-artificial-intelligence-is-a-threat/.
Mann S. Wearable computing: toward humanistic intelligence. IEEE Intell Syst 2001;16(3):10–15, at http://n1nlf-1.eecg.toronto.edu/ieeeis_intro.pdf.
Mann S, Niedzviecki H. Cyborg: digital destiny and human possibility in the age of the wearable computer. Random House; 2001.
Manwaring K, Clarke R. Surfing the third wave of computing: a framework for research into eObjects. Comput Law Secur Rev 2015;31(5):586–603, PrePrint at http://www.rogerclarke.com/II/SSRN-id2613198.pdf.
Marcus G. Deep learning: a critical appraisal, arXiv, 2018, at https://arxiv.org/pdf/1801.00631.pdf.
Markoff J. Scientists See Promise in Deep-Learning Programs. The New York Times; 2012, at https://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html.
Mayer-Schonberger V, Cukier K. Big data: a revolution that will transform how we live, work and think. John Murray; 2013.
Mueller 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, at http://www.informatik.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/MuFr03.pdf.
Paden B, Cap M, Yong SZ, Yershov D, Frazzoli E. A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Trans Intell Veh 2016;1(1), at https://arxiv.org/pdf/1604.07446.pdf.
Robertson LJ, Abbas R, Alici G, Munoz A, Michael K. Engineering-based design methodology for embedding ethics in autonomous robots. Proc. IEEE 2019;107(3):582–99, at https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8620254.
Russell SJ, Norvig P. Artificial intelligence: a modern approach. 2nd edition. Prentice Hall; 2003.
Russell SJ, Norvig P. Artificial intelligence: a modern approach. 3rd ed. Prentice Hall; 2009.
Scherer MU. Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. Harvard J Law Technol 2016;29(2):353–400, at http://euro.ecom.cmu.edu/program/law/08-732/AI/Scherer.pdf.
Selbst AD, Powles J. Meaningful information and the right to explanation. Int Data Priv Law 2017;7(4):233–42, at https://academic.oup.com/idpl/article/7/4/233/4762325.
Sheridan TB, Verplank WL. Human and computer control for undersea teleoperators. MIT Press; 1978, at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.694.7165&rep=rep1&type=pdf.
Simon HA. (1960) The shape of automation reprinted in various forms, 1960, 1965, quoted in Weizenbaum J. (1976), pp. 244–245.
Simon HA. The sciences of the artificial. 3rd ed. MIT Press; 1996.
Simonite T. A Plea For AI That Serves Humanity Instead of Replacing it, 22. Wired Magazine; June 2018, at https://www.wired.com/story/a-plea-for-ai-that-serves-humanity-instead-of-replacing-it/.
Sulleyman A. Elon Musk: AI is a 'fundamental existential risk for human-civilisation' and creators must slow down. Indep 2017, at https://www.independent.co.uk/life-style/gadgets-and-tech/news/elon-musk-ai-human-civilisation-existential-risk-artificial-intelligence-creator-slow-down-tesla-a7845491.html.
Wachter S, Mittelstadt B. 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.
Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 1996;12(4):5–33.
Warwick K. The Cyborg revolution. Nanoethics 2014;8(3):263–73.
Weizenbaum J. Computer power and human reason. W.H. Freeman & Co.; 1976.

Widom J. Research problems in data warehousing. Proceedings of the fourth international conference on informtion & knowledge management; 1995, at http://ilpubs.stanford.edu:8090/91/1/1995-24.pdf.

Wyndham J. The Lost Machine (originally published in 1932), reprinted. In: Wells A, editor. The Best of John Wyndham. London: Sphere Books; 1932 pp. 13–36 and in Asimov I, Warrick PS & Greenberg MH, editors. Machines That Think Holt, Rinehart, and Wilson, 1983, pp. 29–49.

Yampolskiy RV & Spellchecker MS. Artificial intelligence safety and cybersecurity: a timeline of AI failures. arXiv, 2016, at https://arxiv.org/pdf/1610.07997.

Zhaohui W, et al. Cyborg intelligence: recent progress and future directions. IEEE Intell Syst 2016;31(6):44–50.