

Regulatory Alternatives for AI

Review Draft of 9 February 2019

Roger Clarke **

© Xamax Consultancy Pty Ltd, 2019

This document is at <http://www.rogerclarke.com/EC/RAI.html>

Abstract

Artificial Intelligence (AI) is enjoying another of its periodic surges in popularity. To the extent that the current promises are fulfilled, AI may deliver considerable benefits. Whether or not it does so, AI harbours substantial threats. The risks can be managed in various ways, including tolerance, self- and industry-regulation, co-regulatory arrangements and formal law.

Because of the technical and political complexities, and the intensity of the threats, the softer regulatory forms appear to be inadequate. Co-regulation appears to be the most appropriate approach, comprising the establishment of a legislated framework that declares requirements, enforcement processes and sanctions, and allocates powers and responsibilities to appropriate regulatory agencies, but delegates development and maintenance of the detailed obligations to an independent body, comprising representatives of all stakeholder groups, including the various categories of the affected public.

Contents

1. Introduction
2. AI
 - 2.1 A Working Definition for AI
 - 2.2 Four Forms of AI
 - 2.3 The Threats Inherent in AI
 - 2.4 Key Considerations in Regulating AI
3. Regulation
 - 3.1 Regulatory Concepts
 - 3.2 Natural Controls and the Justification of Intervention
 - 3.3 Regulatory Forms
 - 3.4 Regulatory Indicators
4. A Co-Regulatory Framework for AI
5. Conclusions

References

1. Introduction

The term Artificial Intelligence (AI) refers to a suite of technologies whose intention is to exhibit behaviour comparable to, or better than, that of intelligent beings. Current manifestations of AI that are again attracting considerable attention are various kinds of robotics, and a particular form of software development, based loosely on biological neural networks, which uses a set of previous examples to generate a set of weightings, and then applies those weightings to future instances.

AI is claimed to offer great promise, in such ways as labour-savings, more rapid, or reliable, or better-quality, decision-making and action, and the discovery of new and valuable information that would otherwise have remained hidden. On the other hand, AI embodies many and serious threats, in such areas as inscrutable and unaccountable decision-making, hidden error, bias and discrimination, and irrevocable delegation to autonomous artefacts and systems.

A reasoned analysis is needed of appropriate ways in which risks can be managed and hence benefits can be achieved. One form of guidance is the precautionary principle, which assigns moral responsibility to the sponsors of risk-prone initiatives: "When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause and effect relationships are not fully established scientifically" (Wingspread 1998).

This article canvasses the possibilities, takes into account the conditions under which the various approaches are applicable, and concludes with a proposal for a multi-partite framework for the public risk management of AI.

2. AI

As a basis for the analysis that follows, this section provides a working definition of AI, outlines current forms, briefly discusses key threats, and identifies some key considerations involved in designing regulation for AI.

2.1 A Working Definition for AI

Since the term was coined in 1955, AI has been surrounded by an aura of mystery, confusion and a considerable amount of over-claiming. It has accordingly stimulated periods of enormous enthusiasm, interspersed with 'winters of discontent'.

Conventionally in the AI field (Albus 1991, Russell & Norvig 2003, McCarthy 2007):

'intelligence' is exhibited by an artefact if it evidences perception and cognition of (relevant aspects of) its environment, has goals, and formulates actions towards the achievement of those goals

The term 'artificial' implies 'human-made', but the yardstick is open to interpretation variously as 'equivalent to human', 'different from human' or perhaps 'superior to human'.

Given that artefacts and humans have comparative advantages over one another, it can be argued that a more appropriate term would be 'complementary intelligence': "information technologists [should] delineate the relationship between robots and people by applying the concept of decision structuredness to blend computer-based and human elements advantageously" (Clarke 1989, 1993, 2014):

'complementary intelligence' would (1) do things well that humans do badly or cannot do at all; and (2) function as elements within systems that include both humans and artefacts, with effective, efficient and adaptable interfacing among them all

At this stage, however, the world is grinding onwards with the longstanding, but confused and unhelpful, notion of AI.

2.2 Four Key Forms of AI

Some technologies that were once regarded as being AI have made their escape, and forged their own future. Chief among these are various forms of pattern recognition. This section identifies four broad areas of AI that fit within the broad notion. Two of them are currently attracting a great deal of public attention, and appear to be the primary focus of regulatory considerations.

Robotics continues 'on the factory floor' and in warehouses. It provides low-level control over the attitude, position and course of craft on or in water and in the air. It is achieving market penetration in self-driving vehicles, variously on rails and otherwise, in controlled environments such as mines, quarries and dedicated bus routes, but recently also in more open environments. Further examples can be found in the 'Internet of Things' (IoT) movement, and related initiatives under such rubrics as 'smart houses' and 'smart cities'. Robotics, or particular sub-categories of it, may well be a focal point for regulatory activity.

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

A third area of AI that has been in use for some time is **rule-based expert systems**. These involve the representation of human expertise in a form that can

be applied by computers to new data. Rules are statements of relationships between variables. The relationships may be theoretically-based, empirically-derived, mere '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 & Riley 1998). 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 (Clarke 1991).

The fourth form of AI, and the second that appears likely to stimulate regulatory action, is commonly known as '**neural networks**'. This is a branch of the machine learning (ML) area and is based on models whose form is loosely based on biological neural networks. This differs from previous approaches, in that it does not necessarily begin with active and careful modelling of a real-world problem-solution, problem or even problem-domain. Rather than comprising a set of entities and relationships that mirrors the key elements and processes of a real-world system, a neural network model may be simply a list of input variables and a list of output variables (and, in the case of 'deep' networks, intermediary variables). The weightings imputed for each arc within the network reflect the characteristics of the training-set that was fed in, and of the particular learning algorithm that was imposed on the training-set. The implicit claim of proponents of the technique is that its analytical and theoretical weaknesses matter little, and that its almost entirely empirical justification is sufficient. These features combine with questions about the selectivity, accuracy and compatibility of the data to give rise to considerable uncertainty about the technique's degree of affinity with the real world to which it is applied.

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

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

2.3 The Threats Inherent in AI

An expanding literature identifies a wide array of serious concerns about AI's impacts and implications (e.g. Scherer 2016, esp. pp. 362-373, Yampolskiy & Spellchecker 2016, Duursma 2018).

At the heart of the concerns are:

- naive assumptions about:
 - the quality of data
 - the comprehensiveness of available data
 - the compatibility of data drawn from different sources
 - the desirability of ever-more and ever-more granular data
 - the adequacy of empirical correlation unguided by theory and in the absence of rational explanation
- opaque technology, artefacts and systems, embodying:
- errors, and undesirable discrimination, resulting in:
- lack of decision transparency, together leading to:
 - unreplicability
 - unauditability
 - undiscoverability or error and undesirable discrimination
 - uncorrectability
 - unaccountability
 - cavalier behaviour of unaccountable entities

Broader implications of these problems, identified by many authors, include:

- the undermining of human rights
- the unrecallable delegation of autonomy to artefacts that embody models of reality that are weak and insufficiently adaptable
- the disruption of culture
- the disruption of work-based income-distribution
- the imposition of predestination on individuals
- the dominance of collectivism over individualism
- the dominance of the powerful over the weak
- meaninglessness of human life

2.4 Key Considerations in Regulating AI

Artefacts are being imbued with a much greater degree of autonomy than was the case in the past. Autonomy may merely comprise a substantial repertoire of pre-programmed stimulus-response relationships. Alternatively, it may extend to the capacity to adapt aspects of those relationships, or to create new ones. In either case, 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 institutional processes.

Autonomous artefacts also have a high likelihood of stimulating repugnance among a proportion of the public, and hence giving rise to luddite behaviour.

Another consideration is the existence of a succession of phases in the AI supply-chain, from laboratory experiment to deployment in the field. Responsibilities need to be assigned to the various categories of entities that are active in each phase, commensurate with the roles that they play.

In order to deliver AI-based systems, 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	Direct Responsibility
Research	AI Technology	Researchers
Invention	AI-Based Artefacts	IR&D Engineers
Innovation	AI-Based Systems	Developers
Dissemination	Installed AI-Based Systems	Purveyors
Application	Impacts	User Organisations

3. Regulation

This section briefly summarises key information from regulatory theory, in order to establish a basis on which the analysis of alternative approaches to the regulation of AI can proceed. Natural controls are discussed, enabling definition of a threshold test for regulatory intervention. An overview is then provided of the range of forms that regulatory measures can take, and criteria are presented for the design and evaluation of a regulatory regime. Finally, factors are

suggested that determine whether particular regulatory forms are suitable to any particular context.

3.1 Regulatory Concepts

There are many definitions of the notion of 'regulation'. See, for example, Black (2008) and Brownsword & Goodwin (2012). An instrumentally useful definition of regulation is adopted in this work:

***Regulation** is the exercise of control over the behaviours of entities*

This definition is phrased in such a manner as to encompass accidental and incidental control mechanisms rather than only purpose-designed instruments of policy. Not only does the adopted expression avoid any terms relating to 'means', but it also excludes the 'ends' to which the regulation is addressed. The objectives of regulatory schemes are commonly contested, and they change over time. In addition, the effectiveness with which the 'means' achieve the 'ends' is not a definitional feature, but rather an attribute of regulatory regimes.

The entities whose behaviour is subject to control are not the only ones involved in regulation. The following three categories are distinguished:

- **'regulatees'** are entities that are subject to a regulatory scheme. Regulatees may include corporations, unincorporated business enterprises, government agencies, cooperatives, incorporated and unincorporated associations, and individuals
- **'regulators'** are entities that exercise powers in order to achieve control over the behaviours of regulatees. Regulators may include tightly-controlled government agencies and relatively independent commissions, but also incorporated bodies such as stock exchanges, and in some cases industry associations that administer Codes
- **'beneficiaries'** are entities that are advantaged by the regulatory arrangements. The advantages may be intended, accidental or incidental. Encompassing more than only the intended advantages is useful, because most schemes have accidental winners (e.g. incumbent regulatees may gain advantages over new entrants), and those stakeholders naturally become opponents of change to the system. The categories of entities include all of the forms listed under 'regulatees'. However, it can also be useful to broaden the notion to encompass social values such as trust in social and economic institutions, and environmental values

The more comprehensive model in Clarke (2018a) identifies many further actors within the regulatory field. Of particular importance are representatives of and intermediaries for regulatees (such as lawyers, insurers, financiers and consultants) and advocates for the interests of beneficiaries. They support flows of market signals, which are crucial to an effective regulatory scheme.

The design of regulatory regimes reflects the aims of players in the processes of regulation, de-regulation and re-regulation. The coherence and completeness

varies greatly among schemes, depending on the degree of conflict among interests and the power of the parties involved in the design process.

Guidance in relation to the design of regulatory regimes, and in relation to the evaluation of existing schemes, is provided by the criteria listed in Table 2. This was developed by drawing on a wide range of literature, with Gunningham et al. (1998), Hepburn (2006) and ANAO (2007) being particularly useful.

A large body of theory exists relating to regulatory mechanisms (Braithwaite 1982, Braithwaite & Drahos 2000, Drahos 2017). During the second half of the 20th century, a regulatory scheme involved a regulatory body that had available to it a comprehensive, gradated range of measures, in the form an 'enforcement pyramid' or 'compliance pyramid' (Ayres & Braithwaite 1992, p. 35). That model envisages a broad base of encouragement, including education and guidance, which underpins mediation and arbitration, with sanctions and enforcement mechanisms such as directions and restrictions available for use when necessary, and suspension and cancellation powers to deal with serious or repeated breaches.

In recent decades, however, further forms of regulation have emerged, many of them reflecting the power of regulatees to resist and subvert the exercise of power over their behaviour. The notion of 'governance' has been supplanting the notion of 'government', with Parliaments and Governments in many countries withdrawing from the formal regulation of industries (Scott 2004, Jordan et al. 2005). Much recent literature has focussed on deregulation, through such mechanisms as 'regulatory impact assessments' designed to justify the ratcheting down of measures that constrain corporate freedom, and euphemisms such as 'better regulation' to disguise the easing of corporations' 'compliance burden'.

3.2 Natural Controls and the Justification of Intervention

Some controls over behaviour arise from natural influences, by which is meant processes that are intrinsic to the relevant socio-economic system (Clarke 1995, 2014). Examples of **natural regulation** include the exercise of countervailing power by those affected by an initiative, activities by competitors, reputational effects, and cost/benefit trade-offs.

The postulates that an individual who "intends only his own gain" is led by "an invisible hand" to promote the public interest (Smith 1776), and that economic systems are therefore inherently self-regulating, have subsequently been bolstered by transaction cost economics (Williamson 1979). Limits to inherent self-regulation have also been noted, however, such as 'the tragedy of the (unmanaged) commons' notion (Hardin 1968, 1994, Ostrom 1999). Similarly, whereas neo-conservative economists commonly recognise 'market failure' as the sole justification for interventions, Stiglitz (2008) adds 'market irrationality' (e.g. circuit-breakers to stop bandwagon effects in stock markets) and 'distributive justice' (e.g. safety nets and anti-discrimination measures).

Table 2: Criteria for the Design and Evaluation of a Regulatory Regime

Extended version of Clarke & Bennett Moses (2014)

Process

- **Clarity of Aims and Requirements**
Purposes and obligations are understandable by regulatees and beneficiaries
- **Transparency**
Development and review processes are open, and requirements are published
- **Participation**
All stakeholders are involved in development and review processes
- **Reflection of Stakeholder Interests**
The needs of beneficiaries are addressed, and the legitimate interests of regulatees reflected

Product

- **Comprehensiveness**
All relevant aspects are encompassed within a coherent framework
- **Parsimony**
The regime is no more onerous or expensive than is justified
- **Articulation**
The requirements are sufficiently specific and operationalised, to enable effective and efficient implementation by regulatees
- **Educative Value**
Requirements are expressed in explanatory and instructive form, rather than in abstract, discursive prose

Outcomes

- **Oversight**
Regulated behaviours are subject to monitoring
- **Enforceability**
Regulated behaviours are subject to enforcement actions, by beneficiaries directly, and by enforcement agencies
- **Enforcement**
Enforcement agencies have appropriate powers and resources, and apply them in order to achieve compliance
- **Transparency**
Actions taken by regulators, and responses by regulatees, are published, thereby influencing the behaviour of all regulatees
- **Review**
The scheme is reviewed and adapted to ensure that the outcomes correspond to the aims

An appreciation of pre-existing natural controls is a vital precursor to any analysis of regulation, because the starting-point has to be:

'What is there about the natural order of things that is inadequate, and how will intervention improve the situation?

For example, the first of 6 principles proposed by the Australian Productivity Commission was "Governments should not act to address 'problems' through regulation unless a case for action has been clearly established. This should include evaluating and explaining why existing measures are not sufficient to deal with the issue" (PC 2006, p.v). That threshold test is important, in order to ensure a sufficient understanding of the natural controls that exist in the particular context.

In addition, regulatory measures can be designed to reinforce natural controls. For example, approaches that are applicable in a wide variety of contexts include adjusting the cost/benefit/risk balance perceived by the players, by subsidising costs, levying revenues and/or assigning risk.

3.3 Regulatory Forms

In Figure 1, **natural regulation** is depicted as the bottom-most layer of a hierarchy of regulatory forms. Intentionally-designed forms of regulation are depicted as a series of layers, piled on top of the natural form.

The 'instruments' and 'measures' involved represent interventions into natural processes. They are generally designed with the intention to achieve some ends. In principle, the purpose is the curbing of harmful behaviours and excesses, but in some cases the purpose is to give the appearance of doing so, in order to hold off stronger or more effective interventions.

The second-lowest layer in the hierarchy, referred to as **(2) infrastructural regulation**, is a correlate of artefacts like the mechanical steam governor. It comprises particular features of the infrastructure on which the regulatees depend that reinforce positive aspects and inhibit negative aspects of the relevant socio-economic system. Those features may be byproducts of the artefact's design, or they may be retro-fitted onto it, or architected into it. This pattern has been evident since the introduction of steam-engines. Early models did not embody adequate controls over excessive steam-pressure. The first steam-governor was a retro-fitted feature; but, in subsequent iterations, controls became intrinsic to steam-engine design.

Information technology (IT) assists what were previously purely mechanical controls, such as where dam sluice-gate settings are automatically adjusted in response to measures of catchment-area precipitation events or increases in feeder-stream water-flows. IT, and AI-augmented IT, provide many opportunities. One popular expression for infrastructural regulation in the context of IT is 'West Coast Code' (Lessig 1999, Hosein et al. 2003).



Figure 1: A Hierarchy of Regulatory Forms

Switching to the uppermost layer of the regulatory hierarchy, **(7) formal regulation** exercises the power of a parliament through statutes and delegated legislation such as Regulations. In common law countries at least, statutes are supplemented by case law that clarifies the application of the legislation.

A narrow interpretation of law is that it is rules imposed by a politically recognised authority. An expansive interpretation of law, on the other hand, might recognise a much broader set of phenomena, including delegated legislation (such as Regulations); treaties that bind the State; decisions by courts and tribunals, which influence subsequent decisions (depending on the jurisdiction and on court hierarchies); law made by Private Entities, but endorsed and enforced by the State particularly through contracts and enforceable undertakings, and quasi-legal instruments such as memoranda of understanding (MOUs) and formal Guidance Notes (Clarke & Greenleaf 2018).

Formal regulation demands compliance with requirements that are expressed in more or less specific terms, and is complemented by sanctions and enforcement powers. Lessig underlined the distinction between infrastructural and legal measures by referring to formal regulation as 'East Coast code'.

Regulation of the formal kind imposes considerable constraints and costs. The intermediate layers (3)-(6) seek to reduce the considerable constraints and imposts inherent in formal regulation. The lowest of these layers, **(3) organisational self-regulation**, includes internal codes of conduct and 'customer charters', and self-restraint associated with expressions such as 'business ethics' and 'corporate social responsibility' (Parker 2002). However, such mechanisms seldom deliver much advantage to the nominal beneficiaries.

The mid-point of the hierarchy is **(4) industry sector self-regulation**. In many sectors, schemes exist that express technical or process standards. There are also many codes of conduct, or of practice, or of ethics, and some industries feature agreements or Memoranda of Understanding (MoUs) that are claimed to have, and may even have, some regulatory effect. A particular mechanism used in some fields is accreditation ('tick-of-approval' or 'good housekeeping') schemes. These are best understood as meta-brands. The conditions for receiving the tick, and retaining it, are seldom materially protective of the interests of the nominal beneficiaries (Clarke 2001, Moores & Dhillon 2003).

By their nature, and under the influence of trade practices / anti-monopoly / anti-cartel laws, industry self-regulatory mechanisms are generally non-binding and unenforceable. Further, they are subject to gaming by regulatees, in order to reduce their effectiveness and/or onerousness, or to give rise to collateral advantages, such as lock-out of competitors or lock-in of customers. As a result, the two self-regulatory layers are rarely at all effective. Braithwaite (2017) notes that "self-regulation has a formidable history of industry abuse of privilege" (p.124), and the conclusion of Gunningham & Sinclair (2017) is that 'voluntarism' is generally an effective regulatory element only when it exists in combination with 'command-and-control' components.

During the last four decades, several forms have emerged that are intermediate between (often heavy-handed) formal regulation and (mostly ineffective and excusatory) self-regulation.

In Grabowsky (2017), the notion of 'enforced self-regulation' is traced to Braithwaite (1982), and the use of the term '**(6a) meta-regulation**', in its sense of 'government-regulated industry self-regulation', to Gupta & Lad (1983). See also Parker (2007). An example of 'meta-regulation' is the exemption of "media organisations" from Australian data protection law (Privacy Act (Cth) s.7B(4)(b)), provided that "the organisation is publicly committed to observe standards that (i) deal with privacy in the context of the activities of a media organisation (whether or not the standards also deal with other matters); and (ii) have been published in writing by the organisation or a person or body representing a class of media organisations". The absence of any controls is a common feature of such arrangements. The category of meta-regulation is therefore best regarded as pseudo-regulation.

In parallel, the notion of '**(6b) co-regulation**' emerged (Ayres & Braithwaite 1992, Clarke 1999). Broadly, co-regulatory approaches involve enactment of a legislative framework, but expression of the details by means of a negotiation process among the relevant parties. The participants necessarily include at least the regulatory agency, the regulatees and the intended beneficiaries of the regulation, and the process must reflect the needs of all parties, rather than being distorted by institutional and market power. In addition, meaningful sanctions, and enforcement of them, are intrinsic elements of a scheme of this nature.

In contrast with meta-regulation, co-regulation has the scope to deliver effective schemes. However, the promise has seldom been delivered. Commonly, the nominal beneficiaries are effectively excluded from the negotiations, and terms are not meaningfully enforced, and may even be unenforceable (Balleisen & Eisner 2009). Schemes of this kind that lack such fundamentals – typically in the form of 'guidelines' and 'MoUs' but sometimes masquerading under the title of 'Codes' – are referred to in this analysis as **(5) pseudo meta- and co-regulation**.

3.4 Indicators

Each of the regulatory forms identified above may have at least some role to play in any particular context. This section suggests some key factors which variously favour application of the form and militate against its usefulness.

In many circumstances, natural controls can be effective, or at least make significant contributions. They are less likely to be adequate, however, where the social or socio-technical system is complex or obscure, one or a few powerful players dominate the field and can arrange it to suit their own needs, or the field is driven by technologists.

Infrastructural regulation commonly plays a role. IT, including AI-augmentation, provides scope for further improvements in the area, including through the current RegTech movement (Clarke 2018a). It is likely to be at its

least effective, however, in circumstances that involve substantial value-conflicts, variability in context, contingencies, and rapid change.

Organisational self-regulation can only have much effect where the intended beneficiaries have considerable power, or other forms of regulation cause the regulatee to establish protections, and to actually apply them.

Industry self-regulation can only be effective where, on the one hand, a strong industry structure exists, with strong incentives for all industry participants to be members; but, on the other hand, other players are sufficiently powerful to ensure that the scheme delivers advantages to the intended beneficiaries. Unless miscreants feel pain, such schemes lack effectiveness and credibility.

Meta-regulation could only be effective if it imposed a comprehensive set of requirements on the self-regulatory mechanisms. It appears unlikely that many positive exemplars will emerge.

Formal regulation can bring the full force of law to bear. On the other hand, the processes of design, drafting and debate are subject to political forces, and these are in turn influenced by powerful voices, which generally work behind the scene and whose effects are therefore obscured and difficult to detect, far less to counter. As a result, many formal regulatory schemes fail against many of the criteria proposed in Table 2 above. At the other extreme, some formal regulatory arrangements are unduly onerous and expensive, most are inflexible, and all are slow and challenging to adapt to changing circumstances, because of the involvement of powerful voices and political processes.

Co-regulation, on the other hand, offers real prospects of delivering value to beneficiaries. However, there are many obstacles to the establishment of an effective co-regulatory scheme. A trigger is necessary, such as a zealous, powerful and persuasive Minister, or a coalition of interests within or adjacent to a particular sector. A highly-representative forum must come together, and negotiate a workable design. Relevant government(s), government agencies and parliament(s) must have and sustain commitment, and must not succumb to vested interests. A regulator must be empowered and resourced, and supported against the inevitable vicissitudes such schemes encounter.

4. A Co-Regulatory Framework for AI

The previous sections lead to the conclusion that effective regulation can exist through various combinations of natural controls, infrastructural features, co-regulation and formal regulation.

AI comprises multiple technologies, which are embodied in many artefacts, which are embedded into many systems, which are subject to many applications. Some of them feature at least some degree of autonomy. All of them are complex and obscure, and unlikely to be understood even by executives, marketers and policy-makers, let alone by the affected public. AI is dynamic. The entities active

in AI are in many cases small, unstable, rapidly-changing, and short-lived. There is no strong industry structure.

The features of the AI industry militate against natural controls being sufficient, and against infrastructural features being prioritised and implemented. (For example even moderately expensive drones lack communication channel redundancy and collision-detection and -avoidance features). Meanwhile, there are minimal chances of coherent discussions about AI taking place in parliaments, and hence attempts at formal regulation are highly likely to either founder or deliver statutes that are at the same time onerous and ineffective.

Co-regulation, on the other hand, has prospects of enabling the development, implementation and progressive adaptation of effective regulation for AI. The outline provided here could be applied at the level of AI as a whole, although the diversity between the various forms of AI is such that there would be considerable advantages in separating the initiative into technology-specific streams.

At the heart of such a scheme is a comprehensive legislated framework which incorporates at least the following elements:

1. power delegated to an independent Commission or a Minister to approve one or more Codes, and successive versions and replacements of them, subject to:
 - a. a set of requirements with which such Codes must comply
 - b. primacy for negotiated Codes
 - c. a reserve ability to impose Codes if, or to the extent that, negotiated Codes are not achieved
2. one or more Code negotiation and maintenance institutions and processes whose functions are:
 - a. to operationalise the (necessarily abstract) requirements into Codes
 - b. to do so by means of consultative processes
 - c. to achieve active involvement and agreement from all stakeholders including, and especially, the affected public, and
 - d. to reflect the criteria for effective regulation (such as those in Table 2)
3. resources to support that or those institutions and processes
4. enforcement powers and resources, and the obligation to apply those powers and resources
5. the assignment of the enforcement powers and resources to one or more existing and/or new regulatory agencies, whose functions must include overview of consultative processes, supervision of compliance, conduct of own-motion investigations and complaint investigations, imposition of penalties on miscreants, prosecution of offenders, research into technological

and environmental changes, provision of an information clearing house, and provision of a focal point for adaptation of the law, and of Codes

There is scope within such a co-regulatory scheme for various entities to make contributions.

Corporations at all points in the supply-chain can address issues, through intellectual engagement by executives, resource commitment, acculturation of staff, adaptation of business processes, control and audit mechanisms to ensure compliance, establishment, operation and adaptation of internal complaints-handling processes, and communications with other corporations in the supply chain and with other stakeholders, through Code negotiation institutions and processes and otherwise.

Industry Associations can act as focal points for activities within their own sectors. This might include specific guidance for organisations within the particular industry, and second-level complaints processes behind those of the member-corporations; infrastructure that implements protective technologies, and awareness-raising and educational measures.

Individuals need to be empowered, and encouraged, to take appropriate actions on their own behalf. This is only feasible if awareness-raising and educational measures are undertaken, (relatively informal) complaints processes are instituted at corporate and industry association levels, and (relatively formal) complaints, compliance-enforcement and damages-awarding processes are established through regulatory agencies, tribunals and the courts. In some cultures, particularly that of the United States, self-reliance may be writ especially large, while in others, it may play a smaller role, with a correspondingly larger, more powerful and better-funded regulatory agency.

A critical question in the design of one or more regulatory schemes for AI is specifically what is to be regulated. The generic notion of AI is diffuse, and not in itself a suitable target. Regulatory requirements are generally imposed on a category of entity, in respect of a category of activities. An appropriate approach to regulating AI would be to apply the contents of Table 1 above, and impose requirements on entities involved in the research, invention, innovation, dissemination and application of AI technology, artefacts, systems and installed systems.

The forms that relevant regulatory requirements need to take reflect the following modalities (Clarke & Greenleaf 2018):

1. Prohibition– 'You must not'
2. Conditional Prohibition – 'You must not unless'
3. Silence – 'It's up to you'
4. Conditional Permission – 'You may, provided that'
5. Permission – 'You may'
6. Mandation – 'You must'

The question remains as to specifically what technologies and applications within the broad AI field are to be within-scope. An earlier section suggested that two particular forms are the current focus of public attention: Robotics, or some sub-category or categories of it, and data analytics approaches that apply machine-learning techniques such as '(artificial) Neural Networks'.

Feature 1a of the co-regulatory framework described above requires the enunciation of a set of requirements. Recent research has consolidated the following set of principles applicable to AI generally, but with robotics and neural networks particularly in mind:

- 10 General Principles for Responsible AI (Clarke 2018c)
- 50 Specific Principles for Responsible AI (Clarke 2018d)

A related set has been proposed in relation to the specific area of data analytics, with a strong emphasis on neural networks:

- Guidelines for the Responsible Application of Data Analytics (Clarke 2018b)

5. Conclusions

Among the challenges involved in enabling AI to deliver on its promises is the management of the substantial risks that the technologies, artefacts, systems and applications entail. A range of alternative regulatory approaches is feasible. The co-regulatory approach has been argued to be the most appropriate to apply. A degree of articulation of the proposal has been presented.

The considerable amount of public hand-wringing about the risks inherent in AI is of no value unless it stimulates constructive action that addresses those risks. Meanwhile, proponents of promising technologies face the likelihood of strong public and institutional backlash against their innovations. It is therefore in the interests of all stakeholders in AI for a credible public process to be conducted, resulting in a credible regulatory regime that addresses, and is seen to address, the public risks, and that is no more onerous on the AI industry than is justified. The framework proposed in this article provides a blueprint for such a process and such a regime.

References

- 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>
- ANAO (2007) 'Administering Regulation: Better Practice Guide' Australian National Audit Office, March 2007, at http://www.anao.gov.au/~media/Uploads/Documents/administering_regulation_.pdf

- Balleisen E.J. & Eisner M. (2009) 'The Promise and Pitfalls of Co-Regulation: How Governments Can Draw on Private Governance for Public Purpose' Ch. 6 in Moss D. & Cisternino J. (eds.) 'New Perspectives on Regulation' The Tobin Project, 2009, pp.127-149, at http://elearning.muhammadhajirien.org/index.php/catalog/download/filename/New_Perspectives_Full_Text.pdf#page=127
- Black J. (2008) 'Critical Reflections on Regulation' 27 Australian Journal of Legal Philosophy (2002) 1
- Braithwaite J. (1982) 'Enforced self-regulation: A new strategy for corporate crime control' Michigan Law Review 80, 7 (1982) 1466–507
- Braithwaite B. & Drahos P. (2000) 'Global Business Regulation' Cambridge University Press, 2000
- Brownsword R. & Goodwin M. (2012) 'Law in Context: Law and the Technologies of the Twenty-First Century: Text and Materials' Cambridge University Press, 2012
- 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. (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. (1995) 'A Normative Regulatory Framework for Computer Matching' Journal of Computer & Information Law XIII,4 (Summer 1995) 585-633, PrePrint at <http://www.rogerclarke.com/DV/MatchFrame.html#IntrCtls>
- Clarke R. (1999) 'Internet Privacy Concerns Confirm the Case for Intervention' Commun. ACM 42, 2 (February 1999) 60-67, PrePrint at <http://www.rogerclarke.com/DV/CACM99.html>
- Clarke R. (2001) 'Meta-Brands' Privacy Law & Policy Reporter 7, 11 (May 2001), PrePrint at <http://www.rogerclarke.com/DV/MetaBrands.html>
- Clarke R. (2005) 'Human-Artefact Hybridisation: Forms and Consequences' Proc. Ars Electronica 2005 Symposium on Hybrid - Living in Paradox, Linz, Austria, 2-3 September 2005, PrePrint at <http://www.rogerclarke.com/SOS/HAH0505.html>
- Clarke R. (2014) 'What Drones Inherit from Their Ancestors' Computer Law & Security Review 30, 3 (June 2014) 247-262, PrePrint at <http://www.rogerclarke.com/SOS/Drones-I.html>

- Clarke R. (2014) 'The Regulation 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#RN>
- Clarke R. (2018a) 'The Opportunities Afforded by RegTech: A Framework for Regulatory Information Systems' Working Paper, Xamax Consultancy Pty Ltd, April 2018, at <http://www.rogerclarke.com/EC/RTF.html>
- Clarke R. (2018b) 'Guidelines for the Responsible Application of Data Analytics' *Computer Law & Security Review* 34, 3 (May-Jun 2018) 467- 476, <https://doi.org/10.1016/j.clsr.2017.11.002>, PrePrint at <http://www.rogerclarke.com/EC/GDA.html>
- Clarke R. (2018c) 'Principles for Responsible AI' Working Paper, Xamax Consultancy Pty Ltd, October 2018, at <http://www.rogerclarke.com/EC/PRAI.html>
- Clarke R. (2018d) 'Guidelines for the Responsible Business Use of AI' Working Paper, Xamax Consultancy Pty Ltd, October 2018, at <http://www.rogerclarke.com/EC/GAIF.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>
- Clarke R. & Greenleaf G.W. (2018) 'Dataveillance Regulation: A Research Framework' *Journal of Law and Information Science* 25, 1 (2018), PrePrint at <http://www.rogerclarke.com/DV/DVR.html>
- Draho P. (ed.) (2017) 'Regulatory Theory: Foundations and Applications' ANU Press, 2017. at <http://press.anu.edu.au/publications/regulatory-theory/download>
- Duursma (2018) 'The Risks of Artificial Intelligence' Studio OverMorgen, May 2018, at <https://www.jarnoduursma.nl/the-risks-of-artificial-intelligence/>
- Giarratano J.C. & Riley G. (1998) 'Expert Systems' 3rd Ed., PWS Publishing Co. Boston, 1998
- Grabowsky P. (2017) 'Meta-Regulation' Chapter 9 in Draho P. (2017), pp. 149-161, at <http://press-files.anu.edu.au/downloads/press/n2304/pdf/ch09.pdf>
- Gunningham N., Grabosky P, & Sinclair D. (1998) 'Smart Regulation: Designing Environmental Policy' Oxford University Press, 1998
- Gupta, A. & Lad L. (1983) 'Industry self-regulation: An economic, organizational, and political analysis' *The Academy of Management Review* 8, 3 (1983) 416-25
- Hardin G. (1968) 'The Tragedy of the Commons' *Science* 162 (1968) 1243-1248, at <http://cescos.fau.edu/gawliklab/papers/HardinG1968.pdf>
- Hardin (1994) 'Postscript: The tragedy of the unmanaged commons' *Trends in Ecology & Evolution* 9, 5 (May 1994) 199

- Hepburn G. (2006) 'Alternatives To Traditional Regulation' OECD Regulatory Policy Division, undated, apparently of 2006, at <http://www.oecd.org/gov/regulatory-policy/42245468.pdf>
- Hosein G., Tsavios P. & Whitley E. (2003) 'Regulating Architecture and Architectures of Regulation: Contributions from Information Systems' *International Review of Law, Computers and Technology* 17, 1 (2003) 85-98
- Jordan A., Wurzel R.K.W. & Zito A. (2005) 'The Rise of 'New' Policy Instruments in Comparative Perspective: Has Governance Eclipsed Government?' *Political Studies* 53, 3 (September 2005) 477-496
- 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>
- Moore T.T. & Dhillon G. (2003) 'Do privacy seals in e-commerce really work?' *Communications of the ACM* 46, 12 (December 2003) 265-271
- Ostrom E. (1999) 'Coping with Tragedies of the Commons' *Annual Review of Political Science* 2 (June 1999) 493-535, at <https://www.annualreviews.org/doi/full/10.1146/annurev.polisci.2.1.493>
- Parker C. (2002) 'The Open Corporation: Effective Self-regulation and Democracy' Cambridge University Press, 2002
- Parker C. (2007) 'Meta-Regulation: Legal Accountability for Corporate Social Responsibility?' in McBarnet D, Voiculescu A & Campbell T (eds), *The New Corporate Accountability: Corporate Social Responsibility and the Law*, 2007
- PC (2006) 'Rethinking Regulation' Report of the Taskforce on Reducing Regulatory Burdens on Business, Productivity Commission, January 2006, at <http://www.pc.gov.au/research/supporting/regulation-taskforce/report/regulation-taskforce2.pdf>
- Russell S.J. & Norvig P. (2009) 'Artificial Intelligence: A Modern Approach' Prentice Hall, 3rd edition, 2009
- Scott C. (2004) 'Regulation in the Age of Governance: The Rise of the Post-Regulatory State' in J. Jordana J. & Levi-Faur D. (eds) 'The Politics of Regulation' Edward Elgar, 2004
- Scherer M.U. (2016) 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' *Harvard Journal of Law & Technology* 29, 2 (Spring 2016) 354-400
- Smith A. 'The Wealth of Nations' W. Strahan and T. Cadell, London, 1776
- Stiglitz J. (2008) 'Government Failure vs. Market Failure' *Principles of Regulation – Working Paper #144*, Initiative for Policy Dialogue, February 2008, at

http://policydialogue.org/publications/working_papers/government_failure_vs_market_failure/

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

Williamson O.E. (1979) 'Transaction-cost economics: the governance of contractual relations' Journal of Law and Economics 22, 2 (October 1979) 233-261

Wingspread (1998) " Wingspread Statement on the Precautionary Principle, 1998, at <http://sehn.org/wingspread-conference-on-the-precautionary-principle/>

Yampolskiy R.V. & Spellchecker M.S. (2016) 'Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures' arXiv, 2016, at <https://arxiv.org/pdf/1610.07997>

Acknowledgements

This version has benefited from feedback from Prof. Graham Greenleaf of UNSW Law, Sydney.

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 Computer Science at the Australian National University.
