The SECURE framework for Regulating Artificial Intelligence

Tony Carden & Paul Salmon
University of the Sunshine Coast, QLD, Australia

Abstract

AI looks likely to become ubiquitous, offering benefits to a vast variety of human activities. Alongside these benefits will emerge a similarly ubiquitous distribution of risk. These risks run the full range from the inconvenient to the existential. They will manifest in a vast variety of settings and ways. Effectively managing and regulating these risks is critical. The best available risk management and regulatory thinking and practice must be brought to bear on what is perhaps the greatest challenge we have faced as a species. This paper outlines the challenge and proposes a dynamic framework for regulating AI risk underpinned by contemporary regulatory practices that are responsive, smart, agile, interoperative and comprehensive. A regulatory approach based on these principles supports safe, ethical, controllable, usable, responsible, and efficient (SECURE) AI. If implemented, the SECURE framework offers a way to protect the public interest from all classes of AI risks that will scale to all sizes of jurisdiction and support the interoperability of AI regulatory regimes across and between jurisdictions. Further, it will ensure the potential societal benefits of AI are realised.

Introduction

Along with a vast range of benefits, artificial intelligence (AI) presents a wide range of risks. The benefits of AI systems include those already being provided, those currently in development, and those that are aspired toward. These benefits flow from the capacity of AI to undertake cognitive tasks that were previously the exclusive domain of people. AI risks are also created from these same capacities. The need to manage AI risk while optimising its development, deployment and the distribution of its benefits, is one of the great collective challenges of our time.

AI systems are established in a range of domains that already provide known benefits and risks. Benefits of this first tier of widespread AI use are reasonably well known These include enhanced medical diagnosis, rapid vaccine development, autonomous vehicles (land, air, and sea), airport smartgates, improved fraud detection and prevention, logistics, smart manufacturing and content recommendation on social media. Risks of first tier AI are now broadly recognised and many risk controls have been developed and applied in efforts to mitigate them. These risks include bias in training data leading to discriminatory decisions, privacy and security risks, the lack of transparency and explainability, job displacement, risks to civil society through misinformation, and increased wealth and power inequality.

Recent developments and widespread distribution of access to new kinds of AI such as large language and text to image models have introduced a second tier of emergent AI with a vast new set of benefits and risks. Both the benefits and risks of this new, generative AI are, at the time of writing, in their early days. It is notable, however, that many of the risks were largely unforeseen, and many sectors were ill-prepared to manage them. For example, the education sector around the world continues to struggle to adapt to the sudden universal accessibility of tools that can write credible essays and complete many forms of exam on almost any subject. A range of harms are emerging from people misattributing greater skill, generality, intention and even consciousness to emergent AI. In addition, concerns have been raised about generative AI being used to create malware and phishing campaigns, financial fraud, and obscene content involving children.

Many new applications of chatbots like ChatGPT, Bard and LLaMA are being explored and developed in a wide range of areas such as the self-prompting AutoGPT and the task-driven autonomous agents, BabyAGI and SuperAGI. It is critically important to note that, while the foundation models such as GPT-4 that underpin ChatGPT are expensive to build and train, requiring massive compute resources, these latter adaptations are not. This is supporting a wide variety of experimentation with AI system development among a vast array of researchers, small and medium tech businesses, and private individuals. The bar for entry to this kind of experimentation in terms of both knowledge and finances is very low compared with many other risky technologies. This low bar in itself increases risk.

Less widely recognised is the parallel progress in the development of AI systems such as robotics, computer vision, neuroscience-inspired AI, and symbolic AI. In addition to the new benefits and risks that may emerge in each of these discrete areas, the potential combination and permutation of the various strands of AI gives rise to a vast panorama of emergent functionality all of which presents sought-after benefits and unexpected risks. In contrast to the first tier of risk noted above, a far greater degree of uncertainty applies to both the potential and likely benefits and risks of this second tier of AI development and application.

A third tier of AI benefits and risks exist that require qualitatively different approaches to both risk management and risk regulation from either the first or second tiers of AI benefit and risk. The manifestation of these risks is less certain and probably less immediate than tier 1 and 2 risks. They include catastrophic and existential risks posed by more advanced AI systems than those that are evident at the time of writing. This is the domain of human-level and beyond artificial general intelligence (AGI) and artificial super intelligence (ASI), neither of which yet exist. It is speculated and intended by many developers and investors that this level of AI can and will be achieved and will deliver extraordinary benefits such as discovering cures for chronic illness, solving global issues such as disease, environmental damage, climate change, workplace harm, and food and water security, and radically accelerating scientific discovery. Risks of AGI and ASI are expected to emerge from the vast cognitive power of these systems interacting with people and other technologies in ways that could compromise human wellbeing or life on a massive scale.

Given that AI is roughly modelled on organic intelligence – neural networks are rough approximations of brains – the possibility of the emergence of functions other than the current limited cognitive capacities of AI must be considered. For example, consciousness, emotion and creativity that are evident in some organic intelligences may eventually emerge in AI systems. While regulating in advance harms from these systems and functions that do not yet and may never exist is not possible, the level of risk attached to these developments is so high that some preparation and planning is warranted. Given that AGI and ASI, should they emerge, are almost certain to evolve from existing AI, close attention must be paid to precursors of AGI and ASI to mitigate in advance their high-end risks as far as reasonably practicable.

It is important to note that the risk spectrum for AI is vastly different to previous technologies. On top of malicious use or bad design, there are also risks that emerge even when the AI is well designed and works as intended. Furthermore, given that it automates cognitive work, AI is set to permeate society more widely and deeply than any previous technology. This requires a different regulatory approach and critically means we cannot just adopt regulation that has been used previously. In fact, existing regulatory regimes will need to adapt to account for the introduction of AI into the domains they seek to regulate.

Tier 1 risks have a reasonably knowable likelihood of occurrence and consequence. Indeed, some of these risks have already emerged with unacceptable personal, societal, and economic costs. This offers relatively clear pathways for the management and regulation (where regulation approximates assurance of management) of these risks.

The consequences and likelihood of tier 2 risks are inherently less known as they are at any given time, currently emerging. Therefore, a less prescriptive, more flexible, principles-based and agile approach to both the management and regulation of these risks is required.

Tier 3 risks, while less certain and immediate than tier 1 and 2 risks, tend to carry far more severe consequences, up to and including existential risks. A further complication with tier 3 risks is that they are likely only to be actualised by evolutions of AI technology that don't yet exist. Therefore, methods for managing and regulating these risks in advance are far less clear than those for the other two tiers. Notwithstanding this limitation, the likely evolution of current AI into AGI and perhaps ASI may happen in a rapid and non-linear fashion. This necessitates the urgent development and implementation of mechanisms for managing these risks before AGI and ASI is realised. The certainly much slower processes of establishing mechanisms for managing these risks must be done in advance if they are to have any chance of effectiveness. A reactive approach, as seen with tier 1 and 2 risks, will almost certainly result in catastrophic outcomes if relied upon to mitigate tier 3 risks.

It is essential that our best regulatory thinking and practices be brought to bear on this challenge. A key concept that must be included in the design of AI regulation is the difference between probabilistic uncertainty and Boolean uncertainty. Some uncertainties are quantifiable as probabilities, others are not; these latter either will or will not happen. Many current tier 1 risks are quantifiable. For example, there is an in-principle knowable probability that a data set used to train a generative transformer network includes gender or racial bias. In contrast, tier 3 risks such as the possibility that ASI will emerge and compromise the biosphere in pursuit of some self-generated goal is unquantifiable. It either will or will not happen. The attribution of a quantified probability to this risk is illusory, meaningless and misleading. Instead, decisions about how to manage and regulate risks for which the probability of manifestation is unknowable should be guided by the expected severity of consequence.

AI operates within sociotechnical systems. From the small scale of one person interacting with one tool to the massive scale of billions humans interacting with billions of technologies, interconnected through the internet, global travel and trade, it is within the vast interconnected system of nested and dynamic sociotechnical systems that we live our modern lives. Sociotechnical systems are those where people and technologies interact, and where emergent properties arise from these interactions. These are essentially all of our social systems. In sociotechnical systems, technologies afford all kinds of useful but also some harmful processes when people interact with them. For example, a car affords fast and easy personal transport. However, it also affords harmful processes like injury from collisions or environmental damage from emissions. Societies have developed sophisticated and mostly effective ways of optimising the benefits and minimising the risks in sociotechnical systems.

Prior to AI, there was a clear division in sociotechnical systems between technologies and users of those technologies (often labelled as actors or agents). AI blurs the lines as it is both a technology and, increasingly, an agent. While evolving from and continuing as technological elements in those systems, AI is developing and likely to develop more attributes of agents within those systems. Established AI systems already produce effects at the micro and macro scale and everything in between. It's reasonable to expect that future AI systems will permeate all scales of our

sociotechnical systems, offering benefits and presenting risks at all scales. Effective regulation of AI must recognise and account for the formal complexity of the sociotechnical systems of which AI is and will increasingly be a part. This will require the application of complexity theory and sociotechnical systems principles in the design and operation of AI regulatory systems. In practice, this includes recognising that potential harms (risks) emerge, not from any single or 'root' causes, but always from multiple causes. An AI on its own, like any intelligence, cannot cause any harm – or any good. To have any effect, it must interact with other elements in the system – including other humans, agents, and artefacts. This means that opportunities for risk control lie not only in controlling the design and capabilities of AI systems, but also in modifying the other system elements with which AI can interact to have any effect. When risks are high in complex systems, robust risk mitigation requires multiple layers of risk control with intentional redundancy, to cover the inevitable occasional failure of some controls.

How can and should governments respond to AI risk? Our governance, legal and regulatory institutions and mechanisms operate far more slowly than the pace at which AI is developing. How then is it possible to maintain sufficient social control over AI in the public interest? Public interest regulation consists of constraining the actions of private interests to ensure public wellbeing. The familiar applications of this kind of regulation include safety, financial and environmental regulation. The modern history of regulation has seen a shift from the highly prescriptive 'black letter' approaches where governments write and enforce detailed rules, toward progressively more general, principles and outcomes-based regulation. While this trend has helped to reduce red tape and allow expert duty-holders to devise and adapt the details of what compliance looks like, it also raises the prospect of regulatory capture where the vested interests of duty-holders compromise their role in self-regulatory or co-regulatory regimes. Contemporary regulatory methods mitigate these risks by applying systems thinking to the design and operation of regulatory systems and utilising responsive and agile methods.

While all of the three tiers of risk require regulatory oversight, the approach to each tier will differ based on the estimated likelihood and severity of harm and the availability of risk controls. Given the quickly evolving state of AI and the uncertainty about its effects, the regulator should encourage and support innovation in AI risk control. More established tier 1 risks will have known risk controls, emerging tier 2 risks will be subject to experimentation with effective risk control, and prospective tier 3 risks may have few known or effective risk controls. An integrated AI risk regulator must be able to promote, monitor and assure the best available controls for all three tiers of risk.

This paper outlines a framework for AI safety regulation that accounts for the complex nature of the sociotechnical systems in which AI operates, the likely continued rapid evolution of AI capability and functionality and the need to avoid impeding beneficial AI research, development and deployment. The following section outlines some contemporary regulatory principles that underpin current global best practice in regulation. Their application to AI safety regulation is essential.

Regulatory Principles

Responsive regulation

Figure 1. The Responsive Regulatory Pyramid.

Responsive regulation (Ayres & Braithwaite, 1992) is an approach to regulation in which regulators adapt their actions and strategies based on the behaviour, performance, and capacity of duty-holders. Responsive regulators prefer prevention through education and empowerment of duty holders but retain the obligation to monitor compliance and the authority to enforce it where necessary.

The key features of responsive regulation include:

Escalation and de-escalation: Responsive regulation follows a pyramid of regulatory interventions, where regulators start with the least intrusive measures and escalate or de-escalate their actions based on the behaviour of the regulated entities. If a duty-holder complies with regulations, a lighter touch approach is used, while non-compliant entities may face stricter enforcement and more severe sanctions.

Flexibility: Responsive regulation allows for flexibility in the choice of regulatory tools and methods. Regulators can adapt their strategies based on the specific context, industry characteristics, and the nature of the problem being addressed.

Dialogue and cooperation: Responsive regulation encourages open communication and collaboration between regulators and the regulated entities. This fosters trust and mutual understanding, helping to create an environment where compliance is more likely, and problems can be resolved more effectively.

Context sensitivity: Responsive regulation acknowledges that different industries and situations may require different regulatory approaches. By considering the unique context of each case, regulators can design more targeted and effective interventions.

Learning and adaptation: Responsive regulation involves continuous learning and adaptation, as regulators monitor the outcomes of their actions and adjust their strategies accordingly. This ongoing feedback loop helps ensure that regulatory measures adapt and remain effective and relevant over time.

Overall, responsive regulation aims to strike a balance between enforcement and cooperation, using a flexible and context-sensitive approach to encourage compliance and achieve desired regulatory outcomes while minimizing negative impacts on businesses and innovation.

Smart regulation

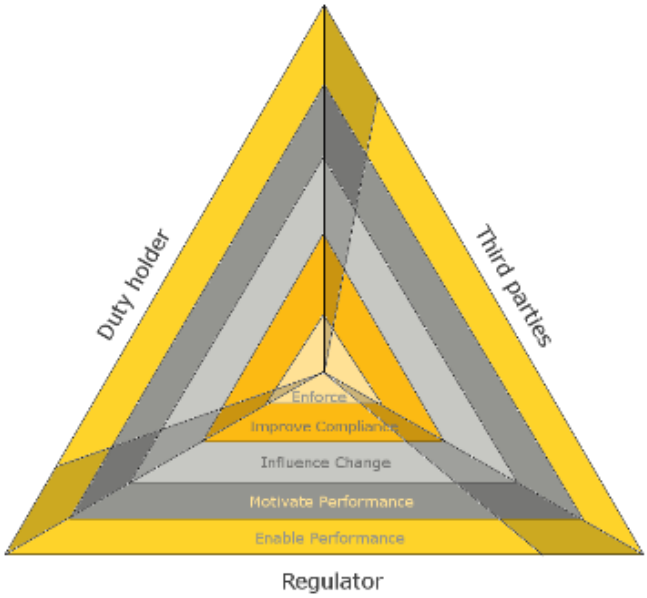

Figure 2. The SMART Regulatory Pyramid.

Smart regulation (Gunningham & Sinclair, 1998) uses the responsive approach described above but applies its principles, first via guided self-regulation (meta-regulation) where duty holders regulate themselves. Where appropriate, third parties are leveraged and incentivised to apply the responsive principles to duty-holders. Given the application of AI across a wide range of domains, a key class of third parties here will be existing regulators. An AI regulator can leverage, educate and support existing regulators to encompass AI safety within their domains. Where these first two lighter-touch avenues are not viable or effective, the regulator itself applies the appropriate tools from the responsive pyramid.

Smart regulation is characterized by the following key elements:

Multiple regulatory tools: Smart regulation recognizes that a single regulatory instrument may not be sufficient or effective in all situations. It advocates for the use of a combination of regulatory tools, such as command-and-control regulations, market-based instruments, self-regulation, voluntary agreements, and information-based strategies, to achieve policy objectives more efficiently.

Co-regulation and partnerships: Smart regulation emphasizes the importance of collaboration between government regulators, businesses, non-governmental organizations, and other stakeholders. By engaging multiple parties in the regulatory process, smart regulation seeks to harness their unique expertise, resources, and perspectives to develop more effective and innovative solutions to regulatory challenges. In the case of AI regulation, key third parties include other regulators.

Flexibility and adaptability: Smart regulation acknowledges that industries, technologies, and environmental challenges evolve over time. To remain effective, regulatory frameworks must be

flexible and adaptable, allowing for adjustments based on new information, changing circumstances, and lessons learned from experience.

Context sensitivity: Smart regulation emphasizes the need to consider the specific context and characteristics of the industry or issue being regulated. This includes understanding the social, economic, and technological factors that may influence the effectiveness of regulatory measures.

Performance-based and outcome-focused: Smart regulation shifts the focus from prescriptive compliance with specific rules to achieving desired outcomes, such as environmental protection or public safety. This performance-based approach encourages innovation and allows regulated entities to develop tailored solutions to meet regulatory objectives.

Monitoring, evaluation, and feedback: Smart regulation involves regular monitoring and evaluation of regulatory measures to assess their effectiveness and identify areas for improvement. This feedback loop ensures that regulatory frameworks remain relevant and responsive to changing needs and conditions.

By integrating these elements, smart regulation seeks to create more efficient, effective, and innovative regulatory frameworks that achieve policy objectives while minimizing the burden on businesses and promoting economic growth.

Agile regulation

Agile regulation (Coglianese, 2023) is a concept referring to a regulatory approach that is flexible, adaptive, and responsive to the fast-paced changes and developments in industries, particularly in technology and innovation. The term draws inspiration from the Agile methodology used in software development, where rapid iterations, continuous feedback, and adaptation are key principles. This approach helps account for the complexity of the sociotechnical system of which AI is a part. Complex systems comprise multiple components that interact dynamically in a non-linear manner, with small events able to produce large and unforeseen outcomes (Cilliers, 1998; Dekker, 2011). Notably in the present context, systems components are unaware of what the entire system comprises, and how their behaviour influences system behaviour. Sociotechnical systems theory provides a set of values and principles for managing technology insertion. A central tenet of this approach is that joint optimisation is required for safe and efficient system performance (Badham, Clegg, & Wall, 2000). This is in contrast to the optimisation of either the social or technical aspects in isolation. Relevant values in the present context include the treatment of humans as assets, that technology should be a tool to assist, not replace humans, respect for individual differences, and responsibility to all stakeholders. An agile AI regulatory system must be agile in keeping pace with the fluid state of the regulated environment, agile in adaptive rulemaking, and agile in moving between responsive regulatory tools.

In the context of regulation, Agile means:

Flexibility: Rather than having rigid rules, Agile regulation allows for adjustments based on the specific needs and circumstances of different industries or businesses. This flexibility enables regulators to adapt to evolving technologies and market conditions.

Collaboration: Agile regulation encourages collaboration between regulators, industry stakeholders, and other relevant parties to foster an environment of shared understanding, learning, and problem-solving. This collaborative approach helps create more informed, effective, and timely regulations.

Iterative and Incremental: Similar to the Agile methodology in software development, Agile regulation adopts an iterative and incremental approach. This means that regulations can be updated and improved as new information becomes available, rather than waiting for major revisions.

Experimentation: Agile regulation promotes the use of pilot programs, sandboxes, and other experimental regulatory frameworks to test and assess the impact of new technologies or business models before implementing full-scale regulations.

Outcome-focused: Agile regulation prioritizes the achievement of desired outcomes, such as consumer protection, fair competition, and public safety, over prescriptive compliance with specific rules. This approach encourages innovation while still ensuring that regulatory objectives are met.

By incorporating these principles, Agile regulation aims to create a more dynamic, effective, and efficient regulatory environment that supports innovation and economic growth while safeguarding the public interest.

An AI regulatory system that is responsive, smart and agile will use the most efficient tools from the responsive pyramid, applied through the most effective of the three smart pathways of monitored self-regulation, third party influence, or direct action by the regulator. Agility will mean that, while the authority, resourcing and principles of the regulatory system should be legislated, the detail of rulemaking should occur further down the system and be articulated through compliance codes and standards that can be rapidly adapted as conditions change,

Regulatory Practice

The overall purpose of AI safety regulation should be to assure that sufficient constraints are in place to support public safety. As in other complex sociotechnical systems, effective AI safety regulation will need to promote monitor and assure a range of risk controls across the system. An example is road safety where risk controls include vehicle safety standards, road infrastructure safety standards, road rules, driver licensing and vehicle roadworthy requirements. Similarly, AI safety requires risk controls to be applied across the system in which AI operates, engaging multiple classes of duty holder such as developers, suppliers, modifiers, and users.

The application of the principles described above will vary per jurisdiction based on the number, mix and type of duty-holders in each jurisdiction. However, in the interests of optimising regulatory effectiveness and efficiency, what should not vary is that the design and operation of each AI regulatory system should be responsive, smart and agile, and should cover all three tiers of AI risk:

1. Known risks associated with established AI systems (current risks)
2. Likely risks associated with emerging AI systems (emerging risks)
3. Possible risks from future AI systems (prospective risks)

Design and operation of an AI regulatory system

In order to operate in a responsive, smart and agile way, an AI regulatory system must be appropriately designed. This requires statutory, policy and procedural foundations that support the responsive, smart and agile approach. Empowering legislation should, therefore, establish powers, limits and principles for regulation but refrain from any detailed rule-making. The detail of how AI risk is managed will need to evolve as AI systems and the broader systems within which they are

embedded evolve. For rule-making and standard-setting, this means continuous, broad consultative processes must be built into the foundations of the regulatory structure.

Any central AI regulatory agency should include governance, management and operational functions that integrate to continuously monitor and assess existing and emerging risk, dynamically choose the optimal tool from the responsive/smart pyramids and apply them with agility. Both governance and rule-making functions of the regulator must include experts from the fields of AI, regulation and law, but also from across academia and society including philosophers, educators, doctors, and community interest groups.

A jurisdiction adopting this approach to AI regulation will need to outline how it will support each of its operational functions in relation to current, emerging and prospective AI risks for each of the 15 focus areas in the smart regulatory pyramid. This will require specifying, establishing and maintaining processes and actions that can enable, motivate, influence, improve and enforce compliance through the agency of (in this order of preference) duty-holders, third parties, and the regulator itself.

Any jurisdiction adopting the framework and principles described here will of course need to adapt them to that jurisdiction with its unique mix of AI researchers, developers, manufacturers, suppliers, and users. The abstraction hierarchy shown in figure 3 represents the functional structure of an AI regulatory system. Actors (individuals, organisations, or AIs) interacting with the objects at the bottom level generate processes shown at the level above. These, in turn, are the means by which the functions and purposes shown at the middle and top levels are achieved. The values in the second top level show what's important and can be used as metrics to monitor system performance.

To enact the principles described in the previous section, an AI regulator could be designed with the functional structure shown in figure 3. In line with smart regulatory principles, the regulator could prioritise the facilitation of the functions shown in figure 3 by duty holders, where that's ineffective by third parties and finally, where that's ineffective, by the regulators own action.

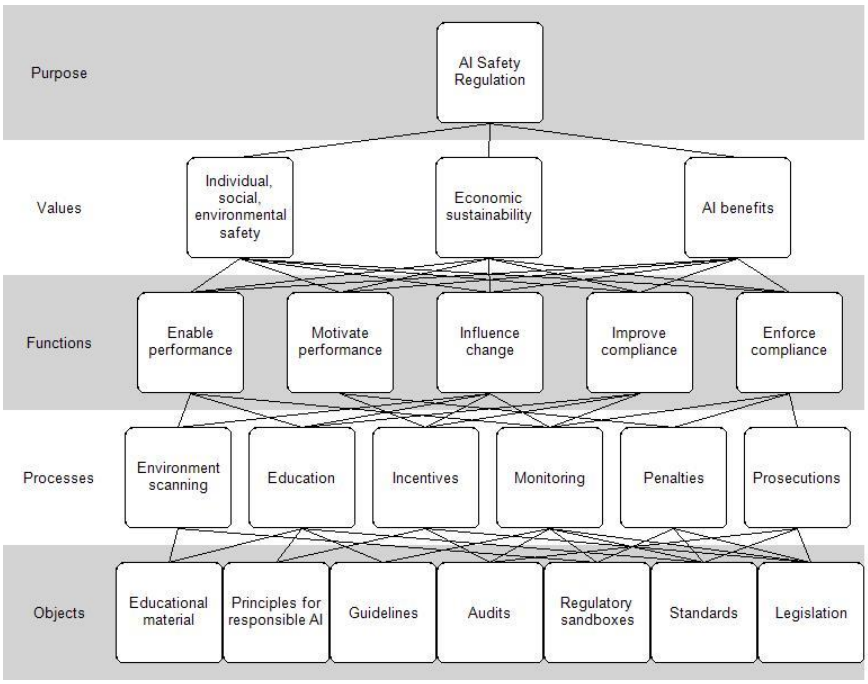

Figure 3. Example functional structure of a responsive, smart, agile AI regulator.

Conclusion

Systems for regulating AI safety are emerging around the world. These include the EU AI Act, the US Blueprint for an AI Bill of Rights, the UK's pro-innovation approach to AI regulation white paper, and China's draft Regulations for Managing Generative AI. While these approaches run the full spectrum from self-regulation (UK) to 'black-letter' hard regulation where government sets and enforces rules (China), none offer a responsive, integrated system that can apply the best regulatory tool for each unique case and context. Furthermore, none cover the full range of potential AI risks represented here by the three tiers. We argue that not only is such an integrated, comprehensive AI regulatory system possible, it is necessary if the public interest is to be adequately protected from AI risks and the full benefits are to be realised.

To be effective, a SECURE system for regulating AI risk must be:

Responsive: by preferring to support compliance with education, influence and incentives while retaining and applying statutory authority to enforce compliance if and where necessary

Smart: by engaging duty holders, other regulators, NGOs and community groups to apply social, economic and reputational pressures toward compliance

Agile: by retaining and applying the capacity for flexibility in rule application, adaptability and new rule making

Jointly optimised: by applying sociotechnical values and principles, accounting for system complexity, and supporting integration of human and technical elements

Comprehensive: an effective AI regulatory system must perpetually and simultaneously address all three tiers of AI risk – current, emerging and prospective.

References:

Ayres, I., & Braithwaite, J. (1992). Responsive Regulation: Transcending the Deregulation Debate. Oxford University Press.

Badham, R., Clegg, C., & Wall, T. (2000). Socio-technical theory. Handbook of Ergonomics. John Wiley.

Coglianese, C. (2023). Regulating machine learning: The challenge of heterogeneity. Competition Policy International: TechReg Chronicle. U of Penn Law School, Public Law Research Paper No. 23-06. https://ssrn.com/abstract=4368604

Gunningham, N., & Sinclair, D. (1998). Smart Regulation: Designing Environmental Policy. Oxford University Press.