NUS Law Working Paper 2021/009

Weapons of Mass Disruption: Artificial Intelligence and International Law

Simon Chesterman

chesterman@nus.edu.sg

[April 2021]

Weapons of Mass Disruption: Artificial Intelligence and International Law

Simon Chesterman[*]

The answers each political community finds to the law reform questions posed artificial intelligence (AI) may differ, but a near-term threat is that AI systems capable of causing harm will not be confined to one jurisdiction — indeed, it may be impossible to link them to a specific jurisdiction at all. This is not a new problem in cybersecurity, though different national approaches to regulation will pose barriers to effective regulation exacerbated by the speed, autonomy, and opacity of AI systems. For that reason, some measure of collective action is needed. Lessons may be learned from efforts to regulate the global commons, as well as moves to outlaw certain products (weapons and drugs, for example) and activities (such as slavery and child sex tourism). The argument advanced here is that regulation, in the sense of public control, requires active involvement of states. To coordinate those activities and enforce global 'red lines', this paper posits a hypothetical International Artificial Intelligence Agency (IAIA), modelled on the agency created after the Second World War to promote peaceful uses of nuclear energy, while deterring or containing its weaponization and other harmful effects.

* Dean and Provost's Chair Professor, National University of Singapore Faculty of Law; Senior Director of AI Governance, AI Singapore. This article was first presented at the 10th Annual Conference of the Cambridge International Law Journal in March 2021. It draws heavily on material discussed at greater length in Simon Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law* (CUP 2021). Many thanks to Denise Cheong, Arif Jamal, Jeong Woo Kim, Nivedita S, Daniel Seng, Alec Stone Sweet, David Tan, Joel Trachtman, Jacob Turner, Ryan Whalen, and Yeong Zee Kin for their comments on earlier versions. Invaluable research assistance was provided by Violet Huang, Eugene Lau, Ong Kye Jing, and Yap Jia Qing. Errors and omissions are due to the author alone.

Around the same time that Isaac Asimov published his short story introducing the three laws of robotics,[1] the world's first nuclear reactor was being built under the viewing stands of a football field at the University of Chicago. There had been some misgivings about initiating a chain reaction in the middle of a densely populated city, but Enrico Fermi, the Italian physicist leading the experiment, calculated that it was safe to do so. On its initial successful run, the Chicago Pile-1 reactor ran for four minutes, generating less than a watt of power — about enough to illuminate one small Christmas tree ornament. The reaction was a major step in the development of nuclear energy, but it was also one of the earliest technical achievements of the Manhattan Project, the US-led initiative during the Second World War culminating in the atomic bombs that incinerated Hiroshima and Nagasaki two and a half years later.[2]

The scientists involved knew that their work had the potential for creation as well as destruction. Though the awesome power of the bomb and the exigencies of war meant that secrecy was an 'unwelcome necessity', Fermi himself believed that preventing the basic knowledge from spreading was akin to hoping the Earth would stop revolving around the Sun.[3] The question was how to ensure that its beneficial use in power generation and medicine did not also lead to proliferation of weapons threatening the existence of humanity.

[1] Isaac Asimov, 'Runaround', *Astounding Science Fiction* (March 1942). These have since become a staple of the literature on regulating new technology though, like the Turing Test, they are more of a cultural touchstone than serious scientific proposal. See Susan Leigh Anderson, 'Asimov's ''Three Laws of Robotics'' and Machine Metaethics' (2008) 22 AI & Society 477.

[2] Richard Rhodes, *The Making of the Atomic Bomb* (Simon & Schuster 1986).

[3] Enrico Fermi, 'Atomic Energy for Power' in AV Hill (ed), *Science and Civilization: The Future of Atomic Energy* (McGraw-Hill 1946) 93 at 103; Enrico Fermi, 'Fermi's Own Story', *Chicago Sun-Times* (23 November 1952).

After the conclusion of the war, that was the subject of the very first resolution passed by the General Assembly of the United Nations in January 1946. It created a commission tasked with recommending how to eliminate such weapons, while enabling all nations to benefit from peaceful uses of nuclear energy.[4] Five months later, the United States, Britain, and Canada proposed that a new international organization be given exclusive control of all aspects of atomic power, from ownership of raw materials to the operation of nuclear power plants. The Soviet Union, wary of Western motives, rejected the plan — a rift that came to be seen as both a cause and a consequence of the Cold War.[5]

It was another seven years before US President Dwight Eisenhower presented an alternative idea to the United Nations. If the earlier plan had been utopian, his 'Atoms for Peace' address was idealistic in a different way: instead of concentrating materials and expertise in a supranational body, they would be disseminated widely — encouraging states to use them for peaceful purposes, in exchange for commitments to renounce the search for the bomb.[6]

The history of efforts to safeguard nuclear power is relevant to the modern challenge of regulating artificial intelligence for three reasons. The first is as an example of a technology with enormous potential for good and ill that has, for the most part, been used positively. Nuclear power, though currently out of favour, is one of few realistic energy alternatives to hydrocarbons; its use in medicine and agriculture is more accepted and widespread. Observers from the dark days of the Cold War anticipated this, but would have been surprised to learn that nuclear weapons were not used in conflict after 1945 and that only a handful of states possess them the better part of a century later.[7]

Secondly, the international regime offers a possible model for regulation of AI at the global level. The grand bargain at the heart of the International Atomic Energy Agency (IAEA), created four years after Eisenhower's speech, was that the beneficial purposes of technology could be distributed in tandem with a mechanism to ensure that those were the only

[4] Establishment of a Commission to Deal with the Problems Raised by the Discovery of Atomic Energy, UN Doc A/RES/1(I) (1946).

[5] Larry G Gerber, 'The Baruch Plan and the Origins of the Cold War' (1982) 6(4) Diplomatic History 69, 70.

[6] Address by Mr. Dwight D Eisenhower, President of the United States of America, to the 470th Plenary Meeting of the United Nations General Assembly (Atoms for Peace) (United Nations, 8 December 1953); Robert L Brown, *Nuclear Authority: The IAEA and the Absolute Weapon* (Georgetown UP 2015) 41-50. By 1953, both Russia and Britain had also conducted successful tests of their own weapons.

[7] For an extreme view, see Kenneth Waltz, The Spread of Nuclear Weapons: More May Better (International Institute for Strategic Studies, Adelphi Papers, Number 171, 1981).

purposes to which it was applied. That trade-off raised the level of trust between the then-superpowers, as well as between the nuclear haves and have-nots. The equivalent weaponization of AI — either narrowly, through the development of autonomous weapon systems, or broadly, in the form of a general AI or superintelligence that might threaten humanity — is today beyond the capacity of most states. For weapon systems at least, that technical gap will not last long.[8] Much as the small number of nuclear armed states is due to the decision of states not to develop such weapons and a non-proliferation regime to verify this, limits on the dangerous application of AI will need to rely on the choices of states as well as enforcement.

A third reason for the comparison is that, much like Fermi and his colleagues, the scientists deeply involved in AI research have been the most vocal in calling for international regulation. The various guides, frameworks, and principles that have been proposed were largely driven by scientists, with states tending to follow rather than lead.[9] As the nuclear non-proliferation regime shows, however, good norms are necessary but not sufficient for effective regulation.

This article considers the institutional possibilities for regulation, with options ranging from a completely free market to global control by an international organization. In between lie more or less formal industry and sectoral associations, as well as public agencies at the national and international level. Rather than laying these out as a menu, a more helpful approach is to focus on the demand for regulation, rather than sources of supply. The management of risks associated with AI can and should, for example, rely heavily on standards that are developed by industry. Best practices, interoperability protocols, and so on will continue to evolve faster than laws can be written. Section one discusses institutional structures that would support rather than hinder that evolution.

Not all risks should be managed, however. It will be necessary to establish red lines to prohibit certain activities. Weaponized or uncontainable AI are the most obvious candidates, but not the only ones. Mere reliance on industry self-restraint will not preserve such prohibitions. Moreover, if those red lines are to be enforced consistently and effectively then some measure of global coordination and cooperation is required. Here the analogy with nuclear weapons is most pertinent. Section two posits a hypothetical International Artificial Intelligence Agency (IAIA), modelled on the IAEA, as a means of achieving this.

[8] See, eg, Elsa B Kania, 'AI Weapons' in China's Military Innovation (Brookings Institution, April 2020).

[9] See Simon Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law* (CUP 2022), ch 7.

The third section returns to the legitimate actions of states. Though the European Union has gone furthest in establishing supranational norms governing the use of AI in the public sector, restrictions on outsourcing of public authority will rely on states themselves for enforcement. Indeed, this will be true of most norms regulating AI. Though industry standards will shape practices and international treaties may limit them, states remain essential players — able to use command and control methods, wielding the 'regulatory hammer', when necessary.[10]

Much as complete internationalization of the nuclear life-cycle in the 1950s was unrealistic and letting the sector develop unchecked was unthinkable, the aim here is to build on existing institutions — most importantly, states — while structuring incentives and coordinating responses. In this way, it should be possible to address these problems of practicality, morality, and legitimacy — ideally, without any bombs going off at all.

1 Industry Standards

The libertarian streak among technology entrepreneurs runs deep. For many years, Bill Gates bragged that Microsoft did not even have an office in Washington, DC — he wanted nothing from the government except to be left alone. Gates was representative of the wider culture in Silicon Valley: most saw their work as undeserving of regulation; a good many deemed themselves morally superior to the governments that might presume to impose it.[11]

In the 2010s this began to change. Three factors appear to have been operating. The first was a growing realization on the part of experts that the potential damage from unchecked innovation did pose a non-trivial risk of catastrophic harm. Much as Fermi and his colleagues saw the dangers of nuclear power, some of the world's leading exponents of technology began to warn of its potential dangers. In addition to public warnings and signing an open

[10] Margot E Kaminski, 'Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability' (2019) 92 Southern California Law Review 1529, 1564. In public international law this is known as the principle of subsidiarity. See Andreas Follesdal, 'The Principle of Subsidiarity as a Constitutional Principle in International Law' (2013) 2 Global Constitutionalism 37.

[11] See, eg, Emanuel Moss and Jacob Metcalf, 'The Ethical Dilemma at the Heart of Big Tech Companies', *Harvard Business Review* (14 November 2019). Cf David Broockman, Greg F Ferenstein, and Neil Malhotra, 'Predispositions and the Political Behavior of American Economic Elites: Evidence from Technology Entrepreneurs' (2019) 63 American Journal of Political Science 212.

letter on the need to ensure that AI remains beneficial, Elon Musk among others donated tens of millions of dollars to the cause.[12]

Secondly, the Cambridge Analytica scandal was a tipping point after which consumer trust in technology companies eroded. The harvesting of data began in 2014 and was used, most prominently, to influence the 2016 US presidential election, but reports had been anonymously sourced until a whistle-blower went on the record in March 2018.[13] Facebook's share price fell by almost a quarter over the following week, losing more than $130bn in market value. Early 2018 was the period in which Microsoft, Google, and IBM all published their AI principles.[14]

A third reason, related to the second, was that companies and researchers correctly anticipated that consumer mistrust would be followed by government action. Though the EU General Data Protection Regulation (GDPR) had been in development for some time, this was also the point that it came into force — even as other jurisdictions were contemplating additional regulation of personal data in particular or technology more generally.[15]

Debates over the obligations of organizations beyond compliance with the law are hardly unique to the technology sector. Linked with larger concerns about the impact of climate change and economic inequality, there is a growing recognition that corporations have responsibilities other than making money.[16] In August 2019, for example, the US Business Roundtable published an open letter on the purpose of a corporation. It stated that its members were committed to delivering value to all their stakeholders: shareholders,

[12] Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter (Future of Life Institute, 2015).

[13] Matthew Rosenberg, Nicholas Confessore, and Carole Cadwalladr, 'How Trump Consultants Exploited the Facebook Data of Millions', *New York Times* (17 March 2018).

[14] The Future Computed: Artificial Intelligence and Its Role in Society (Microsoft, 17 January 2018) 57; IBM's Principles for Trust and Transparency (IBM, 30 May 2018); Artificial Intelligence at Google: Our Principles (Google, 7 June 2018). Facebook announced its own AI ethics team in May of the same year.

[15] These moves were also linked to criticisms concerning the tax strategies and anticompetitive conduct of technology companies.

[16] Cf Simon Chesterman, 'The Turn to Ethics: Disinvestment from Multinational Corporations for Human Rights Violations — The Case of Norway's Sovereign Wealth Fund' (2008) 23 American University International Law Review 577.

employees, suppliers, customers, and communities.[17] The text was unremarkable — such pabulum can be found in annual reports and prospectuses of companies large and small. But to be adopted as policy, signed by 181 chief executive officers of companies from Apple to Walmart, caused a minor stir in economic circles. In particular, it was a public repudiation of the view, championed by Milton Friedman, that the primary responsibility of CEOs is to maximize profits for their shareholders: the business of business, Friedman had argued, is business.[18]

It is not possible in these pages to do justice to the debates over corporate social responsibility or global business activities and human rights.[19] The focus will be on two questions: the role of industry in establishing its own standards for safety, and the limitations of that approach.

1.1 Common Language, Best Practice

One of the most commonly invoked examples of self-governance by researchers is the 1975 Asilomar Conference on recombinant DNA. Given the uncertain dangers associated with the new technique, also known as gene-splicing, US scientists had initially called for a moratorium. The conference brought together more than a hundred biologists from around the world, who developed guidelines for future research. These emphasized the importance of containment as an essential consideration in experiment design, with the level of containment matching, as far as possible, the estimated risk. Certain classes of high-risk experiment for which containment could not be guaranteed were to be 'deferred' — in essence, prohibited.[20] The guidelines were soon endorsed as laws or funding requirements in many countries, with experiments restarting soon afterwards.

[17] Business Roundtable Redefines the Purpose of a Corporation to Promote 'An Economy That Serves All Americans' (Business Roundtable, 19 August 2019).

[18] Milton Friedman, 'The Social Responsibility of Business Is to Increase Its Profits', *New York Times* (13 September 1970). See Claudine Gartenberg and George Serafeim, '181 Top CEOs Have Realized Companies Need a Purpose Beyond Profit', *Harvard Business Review* (20 August 2019).

[19] See generally Abagail McWilliams et al (eds), *The Oxford Handbook of Corporate Social Responsibility: Psychological and Organizational Perspectives* (OUP 2019); John Gerard Ruggie, Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework, UN Doc A/HRC/17/31 (2011).

[20] Paul Berg et al, 'Summary Statement of the Asilomar Conference on Recombinant DNA Molecules' (1975) 72 Proceedings of the National Academy of Sciences 1981.

It is no coincidence that the Future of Life Institute held its own event at the same conference centre some 42 years later to draft the Asilomar AI Principles. Among those principles are an approach to risk that increases control measures commensurate with the expected impact, and an effective prohibition on the development of undirected or uncontainable AI.[21] Yet nostalgia for the 1975 event overestimates the ability of such a gathering to have the same impact today. The biologists involved in the earlier meeting almost all worked at public institutions and were confident that a moratorium would be respected; it was also possible to bring most of the world's leading researchers together at a single event.[22] The disparate and competitive world of AI makes any norms difficult to monitor, let alone police.[23] The Asilomar AI Principles are now merely one of dozens of documents — noticed, to be sure, but hardly authoritative.

Nonetheless, bodies like the Future of Life Institute clearly have a role to play. Apart from anything else, agreeing on terminology can ensure that developers and regulators are not talking past each other. The industry standard to describe 'autonomous' vehicles, for example, follows levels established by the Society of Automotive Engineers (SAE).[24] Similarly, the Institute of Electrical and Electronics Engineers (IEEE) has elaborated principles for ethically aligned design, intended to offer standards and benchmarks for autonomous and intelligent systems.[25]

Indeed, private ordering has governed many aspects of the Internet for decades. Though its origins lie in the US military, since 1998 it has been administered by the Internet Corporation for Assigned Names and Numbers (ICANN), a multi-stakeholder entity with global representation that is incorporated as a non-profit organization in the state of California.[26] This arrangement is desirable because it avoids the problems of either being bound too closely to one state's interests, or held hostage by the lowest common denominator of a

[21] Asilomar AI Principles (Future of Life Institute, 6 January 2017).

[22] Paul Berg, 'Asilomar 1975: DNA Modification Secured' (2008) 455 Nature 290.

[23] 'After Asilomar' (2015) 526 Nature 293. US restrictions on stem cell research in 2001, for example, merely drove research elsewhere.

[24] Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems (revised) (Society of Automotive Engineers, 2018).

[25] Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (IEEE, 2019).

[26] Jeanette Hofmann, Christian Katzenbach, and Kirsten Gollatz, 'Between Coordination and Regulation: Finding the Governance in Internet Governance' (2017) 19 New Media & Society 1406.

group of states.[27] More generally, bodies like the International Organization for Standardization (ISO) establish technical and organizational standards that become de facto norms, despite operating outside traditional structures of domestic or international law.[28] Such standards may be appropriate for emerging industries or practices — among other things, helping to establish what amounts to 'reasonable' conduct for the purposes of determining liability in a claim under tort.

1.2 Perverse Incentives, Regulatory Capture

Standards may be necessary, but they are not sufficient. When working properly, encouraging structured and unstructured conversations among scientists can help build consensus on norms and identify dangerous behaviour, along the lines of the Asilomar recombinant DNA limits. Informal interactions may reveal deviant behaviour, as they did in the case of Russian and South African biological weapons programmes; academic 'gossip' was also instrumental to tracking the Nazi atomic bomb effort during the Second World War.[29] Even if the norms applicable to AI can be agreed on, however, the actors involved in research and development of AI today are too numerous and too diverse to put much hope in industry-wide collective action. A more likely scenario, already apparent in many areas, is fragmentation into regulated and unregulated segments.[30] That is what we see today on the Internet in the form of the dark web.[31]

Alternatively, reliance on self-policing of conduct may lead to organizations seeing regulation as more of a matter of communications than compliance. Much as 'greenwashing' emerged as a method for companies to signal their environmental values without necessarily

[27] Hans Klein, 'ICANN and Internet Governance: Leveraging Technical Coordination to Realize Global Public Policy' (2002) 18 The Information Society 193; Manuel Becker, 'When Public Principals Give Up Control over Private Agents: The New Independence of ICANN in Internet Governance' (2019) 13 Regulation & Governance 561. Cf Jonathan GS Koppell, 'Pathologies of Accountability: ICANN and the Challenge of "Multiple Accountabilities Disorder"' (2005) 65 Public Administration Review 94.

[28] See Nico Krisch and Benedict Kingsbury, 'Global Governance and Global Administrative Law in the International Legal Order' (2006) 17 European Journal of International Law 1.

[29] Jeffery T Richelson, *Spying on the Bomb: American Nuclear Intelligence from Nazi Germany to Iran and North Korea* (Norton 2006) 35.

[30] Stephen M Maurer, *Self-Governance in Science: Community-Based Strategies for Managing Dangerous Knowledge* (CUP 2017) 215-17.

[31] Robert W Gehl, *Weaving the Dark Web: Legitimacy on Freenet, Tor, and I2P* (MIT Press 2018).

committing to specific standards,[32] ethics boards at technology companies have at times been tools of marketing rather than management. Google, for example, launched an Advanced Technology External Advisory Council in March 2019 — then shut it down less than two weeks later due to internal criticism and negative publicity.

Even if standards were universally agreed upon and taken seriously, proximity to industry increases the risk of regulatory capture. This is the phenomenon when those charged with oversight identify more closely with the objectives and problems of the group being regulated, thereby becoming incapable of carrying out their functions independently or effectively.[33] Regulatory capture is not unique to industry regulators — it may apply to government officials, judges, and other actors. Guarding against it is helped by institutionalizing the independence of regulators and reducing the ability to limit the flow of information.[34] Governance at multiple levels can also mitigate the difficulties posed by complexity and the Collingridge Dilemma of *when* to regulate an emerging technology.[35] In the case of AI in particular, connectivity across sectors and borders means that one of those levels needs to be global.

2 Global Red Lines

The effacement of distance is a key structural challenge for regulation of modern technology.[36] Laws in one jurisdiction may not be enforced in another; efforts to prevent or contain deviant behaviour of global reach are only as strong as their weakest link. This is not new and affects various forms of transboundary harm. Willingness to address those deficiencies at the global level has been inconsistent, limited by barriers to agreement due to the nature of international law and impediments to meaningful enforcement for want of powerful institutions. Though international organizations can facilitate the development of

[32] Ho Cheung Brian Lee, Jose M Cruz, and Ramesh Shankar, 'Corporate Social Responsibility (CSR) Issues in Supply Chain Competition: Should Greenwashing Be Regulated?' (2018) 49 Decision Sciences 1088.

[33] Michael E Levine and Jennifer L Forrence, 'Regulatory Capture, Public Interest, and the Public Agenda: Toward a Synthesis' (1990) 6(Special Issue) Journal of Law, Economics, and Organization 167; Jean-Jacques Laffont and Jean Tirole, 'The Politics of Government Decision-Making: A Theory of Regulatory Capture' (1991) 106 Quarterly Journal of Economics 1089.

[34] Ernesto Dal Bó, 'Regulatory Capture: A Review' (2006) 22 Oxford Review of Economic Policy 203.

[35] David Collingridge, *The Social Control of Technology* (Frances Pinter 1980).

[36] See Simon Chesterman, '"Move Fast and Break Things": Law, Technology, and the Problem of Speed' (2021) 33 Singapore Academy of Law Journal 5.

standards, comprehensive global regulation of AI generally is unrealistic and probably undesirable. The focus should therefore be on establishing common red lines for activities that violate fundamental norms or pose significant transboundary threats, with institutional arrangements limited to these purposes.

2.1 Structural Challenges

AI systems are not merely a problem for international organizations to manage; they may undermine such organizations themselves. In part, this is because some AI systems represent a shift of power away from the state. That is true indirectly, through enabling citizens to access information and engage in transactions without the intermediation of traditional public institutions. Yet they may also pose a direct threat to the state, through undermining faith in institutions or processes — spreading 'fake news' and manipulating elections, to pick an extreme but hardly fantastical example.[37]

Historically, international organizations have been ineffective at responding to technological innovation. If regulation lags at the domestic level, it trails internationally.[38] Sovereign equality and the need to reach consensus encourage a lowest common denominator approach to norms, taking years or decades to negotiate. Moreover, the universal membership of forums like the United Nations makes states understandably wary of sharing sensitive information.[39]

Two relevant areas of modest success on the part of international law are banning particular weapons and facilitating global connectivity. From the 1868 St Petersburg Declaration on exploding bullets to more recent attempts to ban landmines and nuclear weapons, international humanitarian law has sought to mitigate human suffering in conflict. This has extended to more recent concerns raised by lethal autonomous weapon systems.[40] International organizations have also supported globalization. One of the oldest such bodies

[37] Eyal Benvenisti, 'Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance?' (2018) 29 European Journal of International Law 9.

[38] Rosemary Rayfuse, 'Public International Law and the Regulation of Emerging Technologies' in Roger Brownsword, Eloise Scotford, and Karen Yeung (eds), *The Oxford Handbook of Law, Regulation, and Technology* (OUP 2017) 500.

[39] Simon Chesterman, 'Does the UN Have Intelligence?' (2006) 48(3) Survival 149.

[40] Simon Chesterman, 'Artificial Intelligence and the Problem of Autonomy' (2020) 1 Notre Dame Journal on Emerging Technologies 210.

is the International Telecommunication Union (ITU), formed in 1865 as the International Telegraph Union before adopting its current name in 1934. Though incorporated as a specialized agency of the United Nations, proposals that it should play a greater role in regulating content on the Internet were met with alarm by many stakeholders — wary that it would restrict the free flow of information online.[41]

The international record is patchier still on providing other public goods. The eradication of smallpox was one of the great achievements of the World Health Organization (WHO), but it took almost two hundred years. A vaccine had been developed in the late eighteenth century, yet it was only after more than a decade of joint global action that the disease was declared eradicated in 1980.[42] As the 2020 Covid-19 pandemic demonstrated, coordinating a global response to a crisis remains extremely difficult when national interests clash.[43] Global action is easiest when the goal is both narrow and shared.[44] In relation to the environment, for example, success in preserving the ozone layer from the damage caused by chlorofluorocarbons may be contrasted with the far greater barriers to addressing global climate change.[45]

Even if there is political will and relative clarity about the activity to be regulated, international law will be ineffective if there is no agreement on the applicable norms, if conduct cannot be attributed to states or other actors at the international level, or if the consequences for breaches are inadequate.

[41] Cf Ramses A Wessel, 'Regulating Technological Innovation Through Informal International Law: The Exercise of International Public Authority by Transnational Actors' in Michiel A Heldeweg and Evisa Kica (eds), *Regulating Technological Innovation* (Palgrave Macmillan 2011) 77; Ingo Take, 'Regulating the Internet Infrastructure: A Comparative Appraisal of the Legitimacy of ICANN, ITU, and the WSIS' (2012) 6 Regulation & Governance 499.

[42] DA Henderson, *Smallpox: The Death of a Disease* (Prometheus 2009).

[43] Peter G Danchin et al, 'The Pandemic Paradox in International Law' (2020) 114 American Journal of International Law 598.

[44] Eyal Benvenisti, 'The WHO — Destined to Fail? Political Cooperation and the Covid-19 Pandemic' (2020) 114 American Journal of International Law 588, 592.

[45] Chris Peloso, 'Crafting an International Climate Change Protocol: Applying the Lessons Learned from the Success of the Montreal Protocol and the Ozone Depletion Problem' (2010) 25 Journal of Land Use & Environmental Law 305.

2.1.1 Norms

On the question of norms, international law generally does not prohibit activities by states unless they have specifically consented to the prohibition.[46] This may take the form of a treaty obligation or customary international law, the latter demonstrated through general practice accepted as law by states.[47] The regime applicable to lethal autonomous weapons, for example, largely draws upon treaties. Treaties are also relevant in establishing human rights norms that prohibit discrimination of the form sometimes perpetuated by AI systems.

Customary international law does regulate certain transboundary harms: states are obliged to ensure that activities within their jurisdiction and control do not cause harm to other states or areas beyond national control.[48] In limited circumstances, this has been expanded by treaty into strict liability. The 1972 Space Liability Convention provides an interesting model whereby a state is 'absolutely liable' to pay compensation for damage caused on the surface of the Earth by space objects launched from its territory.[49] For the most part, however, due diligence is all that is required — based on the nature of the activity, scientific knowledge at the time, and the capabilities of the state in question.[50] As long as this is satisfied, a state will not be responsible for unintentional or accidental acts, including malicious acts by rogue

[46] *Case of the SS 'Lotus' (France v Turkey) (Merits)* (1927 1927) PCIJ Series A, No 10 (Permanent Court of International Justice).

[47] Statute of the International Court of Justice, 1945, art 38(1).

[48] *Legality of the Threat or Use of Nuclear Weapons (Advisory Opinion)* [1996] ICJ Rep 226 (International Court of Justice), para 29. Cf *Corfu Channel (United Kingdom v Albania) (Merits)* [1949] ICJ Rep 4, 22 (every state has an obligation 'not to allow knowingly its territory to be used for acts contrary to the rights of other States').

[49] Convention on International Liability for Damage Caused by Space Objects, done at London, Moscow, and Washington, 29 March 1972, in force 1 September 1972, art II. This may be contrasted with the more limited regime on the high seas where piracy or other hostile activity may serve to absolve a state of its responsibilities. See Joel A Dennerley, 'State Liability for Space Object Collisions: The Proper Interpretation of "Fault" for the Purposes of International Space Law' (2018) 29 European Journal of International Law 281; Trevor Kehrer, 'Closing the Liability Loophole: The Liability Convention and the Future of Conflict in Space' (2019) 20 Chicago Journal of International Law 178. Cf Vienna Convention on Civil Liability for Nuclear Damage, done at Vienna, 21 May 1963, in force 12 November 1977.

[50] *Pulp Mills on the River Uruguay (Argentina v Uruguay) (Judgment)* [2010] ICJ Rep 14, para 197; *Responsibilities and Obligations of States with Respect to Activities in the Area (Advisory Opinion)* [2011] ITLOS Reports 10, paras 117–120.

actors.[51] In such cases, the state's obligation is limited to notification of potentially affected states — though in the case of catastrophic risks that may be insufficient to avert the threat.[52]

In the absence of a treaty, then, the obligations with respect to an AI system that poses transboundary threats — from polluting a river, say, to a general AI capable of seizing military assets — would be due diligence in attempting to prevent the harm and notification if it materializes.

It is important to stress that these obligations fall on states. In areas like human rights, the obligation may be to respect and ensure that rights are protected, sometimes requiring the passage of legislation and administrative action as well as refraining from direct violation of the rights in question.[53] Some international legal obligations do fall directly on individuals — notably the international criminal law regime — but international law first and foremost manages relations between states, only rarely reaching inside them without consent.[54] A key question, then, is whether wrongdoing, or a failure to prevent it, can be attributed to a state.

2.1.2 Attribution

The International Law Commission (ILC) grappled with this topic for half a century, finally producing 'draft' articles on the responsibility of states for internationally wrongful acts that are now accepted as reflecting custom.[55] Completion of the articles was only possible because the ILC deftly set aside the matter of what primary norms might constitute international wrongs to focus on the more technical — and less political — secondary questions of attribution and consequences of liability.

---

[51] Patricia Birnie, Alan Boyle, and Catherine Redgwell, *International Law and the Environment* (3rd edn, Oxford 2009) 147-50.

[52] International Law Commission, Prevention of Transboundary Harm from Hazardous Activities (Articles), UN Doc A/RES/62/68, Annex (2007), art 17; Grant Wilson, 'Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law' (2013) 31 Virginia Environmental Law Journal 307, 342.

[53] Cf Paolo G Carozza, 'Subsidiarity as a Structural Principle of International Human Rights Law' (2003) 97 American Journal of International Law 38.

[54] The most prominent example is enforcement action against a threat to the peace under Chapter VII of the UN Charter. See generally Simon Chesterman, *Just War or Just Peace? Humanitarian Intervention and International Law* (OUP 2001).

[55] James Crawford, *The International Law Commission's Articles on State Responsibility: Introduction, Text and Commentaries* (CUP 2002).

In general, a state is responsible for the acts of 'persons or entities' exercising governmental authority.[56] The term 'governmental authority' is not defined as it depends on 'the particular society, its history and traditions'. Responsibility of the state encompasses situations that involve 'an independent discretion or power to act' on the part of a person or entity — even if the entity 'exceeds its authority or contravenes instructions' while acting in that capacity.[57]

This would cover AI systems used by government agencies and subcontractors, even if the AI system subsequently went beyond intended protocols. The acts of private individuals or corporations would not be covered directly, though the state may have specific treaty commitments or customary obligations to guard against transboundary harm.[58] Failure to satisfy those, at least, is attributable to the state.

Situations may arise where it is difficult to attribute conduct to a particular state or, indeed, to any actor. That is a practical rather than normative challenge, already well known in the context of cybercrime.[59] It points, however, to a potential 'red line' that could be demanded globally: a requirement to ensure that the conduct of AI systems remains traceable back to an entity with a presence in at least one state.

2.1.3 Consequences

The biggest hurdle for international law, however, is the difficulty of enforcing compliance. This is a standard critique of the regime, which suffers from invidious comparisons with domestic legal regimes and periodic accusations that it is not really 'law' at all.[60] The debates are largely sterile due to the dearth of strong theories of international law and the abundance

[56] International Law Commission, Responsibility of States for Internationally Wrongful Acts (Articles on State Responsibility), UN Doc A/56/83, Annex (2001), art 5. The ILC commentary makes clear that 'entity' is not limited to legal persons.

[57] Ibid, art 7; International Law Commission, Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries, UN Doc A/56/10 (2001) 43. Article 8 separately provides that a state is also responsible for the conduct of a 'person or group of persons' if they are in fact acting under the direction or control of that state. The requisite level of 'control' is unclear, but in any case this seems less applicable to truly autonomous AI systems.

[58] See above section 2.1.1.

[59] See, eg, Peter Margulies, 'Sovereignty and Cyber Attacks: Technology's Challenge to the Law of State Responsibility' (2013) 14 Melbourne Journal of International Law 496; Florian J Egloff, 'Public Attribution of Cyber Intrusions' (2021) 6 Journal of Cybersecurity 1.

[60] See, eg, HLA Hart, *The Concept of Law* (2nd edn, Clarendon Press 1994) 213-37.

of practice accepting its legality nonetheless.[61] Those debates fail to take account of structural differences in the normative regimes: international law presumes the horizontal organization of notionally equal sovereign and quasi-sovereign entities, whereas domestic law posits a vertical hierarchy of subjects under a sovereign.[62]

This weakness of international law is a feature, not a bug. Stricter laws would have fewer adherents; more robust institutions fewer members. Nonetheless, mismanaged expectations lead to frustration when collective action problems manifest — as in the case of climate change or pandemics, for example, where international coordination and cooperation are entrusted to institutions lacking the power to impose either.[63]

2.2 An International Artificial Intelligence Agency?

Despite all these caveats, it remains the case that effective regulation of AI requires norms and institutions that operate at the global level. Various scholars and policymakers have recognized this, with the most common prescription being a multi-stakeholder model. Jacob Turner, for example, proposes an analogy with ICANN, the entity that maintains key infrastructure supporting the global Internet.[64] Its elaborate governance model includes representation from the public sector, the private sector, and technical experts. The intuitive appeal is understandable, given the overlap of subject matter and personnel with the AI industry. The actual functions of ICANN are confined to coordinating the Domain Name System and resolving disputes, however.[65] This is important, but the need for a global body to regulate AI goes beyond technical coordination.

In December 2018 Canada and France announced plans to establish an International Panel on AI, modelled on the Intergovernmental Panel on Climate Change (IPCC) established some 30

[61] Louis Henkin, *How Nations Behave: Law and Foreign Policy* (2nd edn, Columbia UP 1979).

[62] See Simon Chesterman, 'An International Rule of Law?' (2008) 56 American Journal of Comparative Law 331.

[63] Sam Johnston, 'The Practice of UN Treaty-Making Concerning Science' in Simon Chesterman, David M Malone, and Santiago Villalpando (eds), *The Oxford Handbook of United Nations Treaties* (OUP 2019) 321 at 328-31.

[64] Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan 2019) 240-42. See also above n 26.

[65] Bylaws for Internet Corporation for Assigned Names and Numbers (ICANN, 1 October 2016) s 1.1.

years earlier.[66] It was later renamed the Global Partnership on AI (GPAI) with a secretariat at the OECD in Paris.[67] The analogy with climate change acknowledges that AI poses a similar collective action problem for the global system. Yet the link with the OECD and an emphasis on human rights point less to concerns about efficient management than a desire to exclude China — indeed, the United States had refused to join due to the potential impact on business but reversed course, citing the need to check China's approach to AI.[68] Experts will take part in working groups on themes including responsible AI, data governance, the future of work, and innovation and commercialization. Worthy goals, but increasing the risk of a bifurcated Internet and approach to AI — the antithesis of a global response.

These and other examples recognize the need for action but also wariness about the practicality and desirability of seeking consensus among states. In theory, for example, the United Nations or the ITU could be entrusted with such a role. They might be helpful forums for norm-setting, but an operational role would inspire reactions comparable to when ITU was proposed as a successor to ICANN to administer the Internet.[69]

International institution-building is an architecture of compromise.[70] Proposals to start with a less formal organization, laying foundations for more elaborate possibilities, reflect the practical challenges of finding common ground.[71] Yet these less ambitious or more political proposals lack both the normative teeth and the aspiration to universalism — the depth and breadth necessary to address the global challenge.

---

[66] France and Canada Create New Expert International Panel on Artificial Intelligence (Gouvernement, 7 December 2018).

[67] Joint Statement From Founding Members of the Global Partnership on Artificial Intelligence (US State Department, 15 June 2020).

[68] Max Chafkin, 'US Will Join G-7 AI Pact, Citing Threat From China', *Bloomberg* (28 May 2020).

[69] See above n 41. In May 2020, the UN Secretary-General produced a report on digital cooperation, identifying key gaps as being a lack of inclusiveness, coordination, and capacity-building. Report of the Secretary-General on the Road Map for Digital Cooperation, UN Doc A/74/821 (2020), para 56.

[70] Timothy LH McCormack and Gerry J Simpson, 'A New International Criminal Law Regime?' (1995) 42 Netherlands International Law Review 177.

[71] See, eg, Olivia J Erdélyi and Judy Goldsmith, 'Regulating Artificial Intelligence: Proposal for a Global Solution' (2018) AAAI/ACM Conference on AI, Ethics, and Society (AIES'18) 95; Jiabao Wang et al, 'Artificial Intelligence and International Norms' in Donghan Jin (ed), *Reconstructing Our Orders: Artificial Intelligence and Human Society* (Springer 2018) 195.

Here the IAEA offers a better model as an example of a regime that confronted a regulatory deficit directly — how to limit the proliferation of nuclear weapons — and embraced the politics of the situation openly: seeking buy-in from non-nuclear states by allowing access to technology, while giving nuclear states assurances that their military advantage would not be lost (at least not until some unspecified point in the future).

As indicated earlier, the IAEA was created at a time of high — perhaps excessive — optimism concerning the potential for nuclear energy, tempered by fears of its weaponization. The Agency's stated objectives are to 'accelerate and enlarge the contribution of atomic energy to peace, health and prosperity throughout the world', while ensuring that this does not further any military purpose.[72] The first of these objectives was pursued through technology transfer, although dreams of electricity 'too cheap to meter' never materialized and more was achieved in medicine and agriculture than power generation.[73] The second objective eventually saw the signing of the Nuclear Non-Proliferation Treaty (NPT). That formalized the two-tier system of nuclear haves and have-nots, with the IAEA tasked with verifying that non-nuclear powers do not divert nuclear material to weapons programmes.[74] The nuclear powers, for their part, committed to 'pursue negotiations in good faith'[75] toward disarmament, but even its own history acknowledges that the IAEA was 'essentially irrelevant' to the nuclear arms race in the course of the Cold War.[76]

Broader standard-setting was, initially at least, an incidental role for the IAEA. Its Statute provides that it can establish 'standards of safety for protection of health and minimization of danger to life and property'.[77] Though the standards are not binding, in practice they are relied upon by states developing and implementing national legislation and standards for nuclear energy.[78] The 1986 Chernobyl disaster revealed major deficiencies in this

[72] Statute of the International Atomic Energy Agency, done at New York, 23 October 1956, in force 29 July 1957, art II.

[73] Brown (n 6) 55-61.

[74] Treaty on the Non-Proliferation of Nuclear Weapons, done at Washington, London, and Moscow, 1 July 1968, in force 5 March 1970, art III.

[75] Ibid, art VI.

[76] David Fischer, *History of the International Atomic Energy Agency: The First Forty Years* (IAEA 1997) 10.

[77] IAEA Statute, art III(A)(6); Paul C Szasz, *The Law and Practices of the International Atomic Energy Agency* (International Atomic Energy Agency 1970).

[78] Philippe Sands and Jacqueline Peel, *Principles of Environmental Law* (4th edn, CUP 2018) 595.

arrangement. A review group recommended better exchanges of information, additional safety standards and guidelines, and enhancing the capacity to perform evaluations. Additional treaties were also concluded, hardening soft law into rules.[79]

A hypothetical IAIA could draw upon the experience of its nuclear counterpart in three ways: the bargain to encourage buy-in, the scope of its authority, and the structure of the organization itself.

2.2.1 Bargain

First, an explicit bargain could bridge the medium-term interests of the most technologically advanced states — the United States and China, for example — and the shorter-term needs of others. The IAEA and the non-proliferation regime were negotiated at a time when the nuclear powers enjoyed a monopoly over its destructive power that they knew could not last. Those states with the most advanced lethal autonomous weapon systems today may come to see that a world in which such weapons are widely distributed would be deeply unstable; if or when advances towards general AI indicate the dangers of a superintelligence, hopes that the technology could be kept secret recall Fermi's warning that the Earth will not cease its motion around the Sun.

Though the link does not appear to have been made before now, the rhetoric of 'AI for Good', used by ITU at its global AI conferences since 2017, has echoes of Eisenhower's 'Atoms for Peace' from 64 years earlier.[80] Where Eisenhower spoke of nuclear energy's potential to be a 'great boon, for the benefit of all mankind', the AI for Good summits emphasize that AI innovation will be central to the achievement of the UN Sustainable Development Goals.[81] Eisenhower's proposal, it should be noted, took time to be accepted by the Soviet Union and was denounced as 'insane' by US Senator Douglas McCarthy.[82] The creation and relative success of the IAEA were tied to the demand for international cooperation on peaceful

[79] Convention on Early Notification of a Nuclear Accident, done at Vienna, 26 September 1986, in force 27 October 1986; Convention on Assistance in the Case of a Nuclear Accident or Radiological Emergency, done at Vienna, 26 September 1986, in force 26 February 1987; Convention on Nuclear Safety, 17 June 1994, in force 24 October 1996; Joint Convention on the Safety of Spent Fuel Management and on the Safety of Radioactive Waste Management, done at Vienna, 5 September 1997, in force 18 June 2001.

[80] See above n 6.

[81] See, eg, AI for Good Global Summit 2017 (ITU, 7-9 June 2017).

[82] 'McCarthy Scorches Plan of Giving Atom Materials', *The News-Review* (Roseburg, OR, 9 February 1957).

nuclear technology and non-proliferation, as well as the clear and delimited role for the new organization.[83]

It is, of course, far from clear that similar conditions obtain today, at a time when the legitimacy of global public institutions has been called into question and the United States and China are, for distinct reasons, especially wary of constraint by external bodies.[84] How to manage the privileges of the powerful without compromising the legitimacy of the organization is one of the trickiest aspects of international institution-building. Acceptance as a nuclear power stands alongside the veto power in the UN Security Council as the most blatant concessions of special privileges based on military might. There is no direct comparison in the field of AI at this point, but an alternative analogy can be drawn with pandemics. After the eventual eradication of smallpox in 1980, all known stocks of the virus were destroyed — with two exceptions. The United States and Russia kept small quantities of the virus: officially because these were the two WHO reference laboratories with the highest security storage facilities; unofficially in deference to the political realities of the Cold War.[85]

2.2.2 Authority

A second lesson from the IAEA is to have a clear and limited normative agenda, with a graduated approach to enforcement. The main 'red line' proposed here would be the weaponization of AI — understood narrowly as the development of lethal autonomous weapon systems lacking 'meaningful human control' and more broadly as the development of AI systems posing a real risk of being uncontrollable or uncontainable.[86]

On the narrower interpretation, it may be asked whether states would ever willingly give up weapons that might provide a military advantage. Yet, in addition to the limits on nuclear weapons, that is precisely what states have done in respect of chemical and biological

[83] Brown (n 6) 64-65.

[84] See Simon Chesterman, 'Can International Law Survive a Rising China?' (2020) 31 European Journal of International Law 1507.

[85] Resolution WHA33.4 (World Health Assembly, 1980), recommendations 9 and 10; Smallpox Eradication: Destruction of Variola Virus Stocks (World Health Organization, A52/5, 15 April 1999).

[86] Cf Draft Report with Recommendations to the Commission on a Civil liability regime for Artificial Intelligence (EU Parliament Committee on Legal Affairs, 2020/2014(INL), 27 April 2020) (distinguishing between 'high-risk' and other applications of AI).

weapons, as well as more recent limitations on blinding weapons.[87] Provided that it could be imposed in a reciprocal manner, there is no reason why a ban on lethal autonomous weapon systems should be unattainable. Indeed, much of international humanitarian law consists of rules that constrain the methods a state may use in armed conflict — accepted because it is known that similar constraints apply to one's potential opponents. Though it is a relatively recent addition, a central justification today is that such laws 'maintain some humanity in warfare'.[88]

The broader interpretation — linked with the question of superintelligence — is more open to debate. There is widespread agreement that AI systems should remain under human control. At present there does not appear to be an immediate danger that an uncontrollable AI in the sense of a sentient being will be created anytime soon. There are, however, many examples of computer viruses that have gotten out of control.[89] The most realistic prospect here would be that states agree to the principle of control, with periodic reviews on progress towards general AI and an accompanying reconsideration of whether limitations on further research are required.[90]

Much as the IAEA developed safety standards over time, these could be an additional function of the proposed IAIA. Standards might draw upon the various principles that have been adopted, but the priority should be human control and transparency. The control aspect applies to autonomous weapons and general AI discussed above. Transparency raises questions that distinct political systems will answer in their own way. In terms of a red line at the international level, however, it would be to require that states prevent AI systems being deployed in a manner that cannot be traced back to a legal person identifiable as the owner,

[87] Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction, done at Washington, London, and Moscow, 10 April 1972, in force 26 March 1975; Convention on the Prohibition of the Development, Production, Stockpiling, and Use of Chemical Weapons and on Their Destruction, done at Paris, 13 January 1993, in force 29 April 1997.

[88] Robert Kolb, 'The Protection of the Individual in Times of War and Peace' in Bardo Fassbender and Anne Peters (eds), *The Oxford Handbook of the History of International Law* (OUP 2012) 317 at 321.

[89] See, eg, Danny Palmer, 'MyDoom: The 15-Year-Old Malware That's Still Being Used in Phishing Attacks in 2019', *Wired* (26 July 2019).

[90] Cf Stephan Guttinger, 'Trust in Science: CRISPR-Cas9 and the Ban on Human Germline Editing' (2018) 24 Science Engineering Ethics 1077.

operator, or manufacturer.[91] The IEEE, for example, stresses the importance of traceability of errors, comparing it to the role of flight data recorders in the field of aviation.[92] The analogy is important with respect to analysing failures, but an even more important equivalent technology is the use of transponders to track aircraft and identify them in the first place.

Such a requirement would not be new to AI. The European Union, for example, requires that high-frequency trading algorithms identify themselves; there is also a growing recognition that AI systems should not pretend that they are human — or be required to make clear that they are not. Proposals to maintain a register of autonomous agents have been floated in the past, drawing upon existing practices such as maintaining a national register of companies.[93] Indeed, in September 2020 Helsinki and Amsterdam launched AI registers as 'a window' to the systems that the cities use.[94] This was laudable as a form of disclosure by public bodies, but given the likely proliferation and pervasiveness of AI systems, registers are unworkable at scale as they would potentially require every computer program to be 'registered'. It might be possible to automate aspects of this, for example mediating transactions through a distributed-ledger regime.[95] AI systems could be required to identify themselves either actively, through notification, or passively, through including a digital signature in their code with a prohibition against removal.

No regime will be perfect or immune to gaming by sophisticated actors. It would need to be supplemented by a forensic capability to identify those responsible for 'rogue' AI systems.

[91] This would include ships at sea (such as those mooted by Google more than a decade ago), which remain under the jurisdiction of a territorially-bounded state. See Steven R Swanson, 'Google Sets Sail: Ocean-Based Server Farms and International Law' (2011) 43 Connecticut Law Review 709.

[92] Ethically Aligned Design (n 25) 137.

[93] See, eg, Curtis EA Karnow, 'Liability for Distributed Artificial Intelligences' (1996) 11 Berkeley Technology Law Journal 147, 193-96; European Parliament Resolution with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)) (European Parliament, 16 February 2017), paras 2, 59. Proposals to establish an IAEA database of nuclear materials were resisted by states due to concerns about compromising commercial information — or the possibility that they might be held responsible in the event that their materials were used in a terrorist incident. Brown (n 6) 162. Instead, states are encouraged to maintain their own national register of sources. Fischer (n 76) 204.

[94] Sarah Wray, 'Helsinki and Amsterdam Launch AI Registers to Detail City Systems', *ITU News Magazine* (30 September 2020).

[95] Cf Turner (n 64) 197-201; Kelvin Low and Eliza Mik, 'Pause the Blockchain Legal Revolution' (2020) 69 International and Comparative Law Quarterly 135.

This would be a challenging — perhaps impossible — task.[96] But the IAIA could serve as a clearinghouse to gather and share information about such systems. Again, a parallel can be found in the IAEA, which established an illicit trafficking database in 1995 to facilitate tracing of nuclear material 'out of regulatory control'.[97]

A final role of the IAIA could be in response to emergencies. Though states would remain the primary actors, it could serve as a focal point for notification of emergencies threatening transboundary harm and coordination of a response. There should be no illusion that a state will be forthcoming in raising the alarm about an uncontrollable or uncontainable AI, particularly if there is a chance that it will not be identified as the source. Indeed, this was Russia's initial response to the Chernobyl nuclear disaster in 1986. States subsequently adopted a treaty obliging parties to notify the IAEA or affected states of any accident within their jurisdiction or control in which release of radioactive material is likely and may be of 'radiological safety significance'.[98] If similar obligations are not accepted by states before the first true AI emergency, they would likely be adopted soon after it.

2.2.3 Structure

A third learning point from the IAEA is the mundane yet important question of structure. Most international organizations are weak by design, with governance powers held closely by member states while management is carried out by a secretariat. The United Nations is the clearest example of this, headed by a Secretary-General whose position in the organization's founding document is styled as its 'Chief Administrative Officer'.[99] The UN Security Council, for its part, is an outlier — a body with real teeth in the form of enforcement powers ranging from economic sanctions to the use of military force. The Council's remit is limited to threats to international peace and security, however, and its powers are firmly under the control of member states. An AI emergency could rise to the level that it justifies Security Council action.

---

[96] Cf Edwin Dauber et al, 'Git Blame Who? Stylistic Authorship Attribution of Small, Incomplete Source Code Fragments' (2017) 1701.05681v3 arXiv.

[97] IAEA Incident and Trafficking Database (ITDB) (IAEA, 2020); Klaus Mayer, Maria Wallenius, and Ian Ray, 'Tracing the Origin of Diverted or Stolen Nuclear Material through Nuclear Forensic Investigations' in Rudolf Avenhaus et al (eds), *Verifying Treaty Compliance: Limiting Weapons of Mass Destruction and Monitoring Kyoto Protocol Provisions* (Springer 2006) 389 at 402.

[98] Convention on Early Notification of a Nuclear Accident, arts 1, 2. See also IAEA Response and Assistance Network (IAEA, 2018).

[99] Charter of the United Nations, done at San Francisco, 26 June 1945, in force 24 October 1945, art 97. See Simon Chesterman (ed), *Secretary or General? The UN Secretary-General in World Politics* (CUP 2007).

Even then, the Council has in the past relied on expert agencies. In the context of counter-proliferation, for example, the Council has drawn on IAEA expertise and resources in relation to North Korea, Iraq, and Iran.

Unusually among intergovernmental organizations, it is the Board of Governors of the IAEA, a subset of member states that meets five times a year — not the General Conference of all members that gathers annually — that has ongoing oversight of its operations, appoints its executive head, and evaluates compliance with its Statute.[100] This has allowed the IAEA to function more effectively, but demands more of the men and women sent as national representatives. Indeed, its history reflects a shift from heads of nuclear agencies in the early years, evangelizing nuclear power, to diplomats more concerned with non-proliferation and budgets.[101]

In the case of a notional IAIA, positioning it as an expert body with additional mechanisms to involve industry, academia, and activists would enhance its legitimacy and relevance. Yet to have 'teeth', it would need to be grounded in the public authority of states.

3 State Responsibility

Tasked with promoting the safe, secure, and peaceful use of nuclear technology, the IAEA is, in the scheme of things, small. With a budget of US$700m and around 2,500 staff, it is comparable in size to the local government of a small town and less than a quarter of the size of Tokyo's Fire Department. Lacking its own enforcement powers, it has relied *in extremis* on the UN Security Council. But compliance — as with most of international law — depends on the behaviour and attitudes of its member states.

As I have argued elsewhere, existing state institutions and norms are capable of regulating most applications of AI.[102] Legislatures, executives, and judiciaries within virtually all states can adapt to fast, autonomous, and opaque AI systems. The effectiveness of those adaptations is tied to the unique legitimacy of public institutions at the state level, which requires that these powers be exercised by officials that are publicly accountable — and not

[100] Brown (n 6) 55. Cf Simon Chesterman, 'Executive Heads' in Jacob Katz Cogan, Ian Hurd, and Ian Johnstone (eds), *The Oxford Handbook of International Organizations* (OUP 2016) 822 at 824.

[101] Fischer (n 76) 425.

[102] See Chesterman (n 9).

themselves outsourced to machines. This section will briefly discuss the roles and the limits of the different branches of government. To identify and fill gaps in the regulatory ecosystem, an independent agency or official with a wide mandate would be an important addition. The example proffered here is an AI Ombudsperson.

3.1 Legislature

Though legislatures around the world have been wary of over-regulating AI systems, they are being forced to enact or amend laws to address anachronisms like presuming that all vehicles have a 'driver'. In addition to ensuring that laws are not skirted because of the speed, autonomy, and opacity of AI systems, additional new laws may be required to ensure human control and transparency.

Legislatures have the advantage of democratic legitimacy, with many jurisdictions favouring them as the body to take decisions on fundamental social policies or involving choices between contested values. Decisions are made by men and women chosen as political representatives rather than subject matter experts, but they have the force of law and are of general application. Because of this, legislatures may be slow to deliberate and their edicts hard to undo. This poses a dilemma for states uncertain about the risks associated with new technology, but also wary of unnecessarily constraining innovation. When there is consensus on the need for clear rules and strong enforcement, legislation is the most legitimate and effective path. Until that time, states may prefer 'masterly inactivity'.

3.2 Executive

Implementation of laws falls to the executive. Agencies tasked with this may develop subject matter expertise and be more flexible in their approach to regulation. In terms of expertise, however, the public sector struggles to keep up with the private sector. This is true in both securities regulation and competition or antitrust law, as well as technology regulation more generally.[103] Flexibility and the ability to react quickly raise accountability questions the further that agencies get from democratic legitimacy. The problem may manifest in over- or

[103] See Chesterman (n 36).

under-zealous regulation, along with the possibility of capture. These concerns can be mitigated through monitoring and review strategies.[104]

Around the world, licensing bodies, product safety regulators, securities regulators, transportation authorities, police forces, national security agencies, and data protection authorities will be at the front line of whether and how to regulate AI systems. Where laws are framed widely or vaguely, significant discretion devolves to these entities. Their ability to act in advance of problems, to publish guidance material, to engage proactively with developers and manufacturers as well as consumers, distinguishes them from other arms of government. When they fail to act, uncertainty may impose its own costs if companies shy away from risky behaviour or if those risks are pushed onto consumers.[105]

3.3 Judiciary

Where harm results or disputes arise, courts may be asked to step in. The strength and the weakness of judicial law-making is its responsiveness to changing circumstances. This enables judges to exercise a modicum of creativity in interpreting the law or applying precedent, but it also means that they are beholden to the cases that come before them. In most jurisdictions, courts are unable to opine on hypothetical situations; when they do so in the common law tradition, their observations are *obiter dicta* — things said in passing that do not bind other tribunals. The *ex post* role of courts may also be a long time *post*: appellate proceedings can take years, meaning a final determination is made only after the technology in question is obsolete.[106]

'Hard cases make bad law', as Oliver Wendell Holmes, Jr, famously warned a century ago. Yet the context from which the cliché is typically lifted adds nuance to this observation. Because hard cases are frequently great ones:

> Great cases like hard cases make bad law. For great cases are called great, not by reason of their real importance in shaping the law of the future, but because of some accident of immediate overwhelming interest which appeals to the feelings and distorts the judgment. These immediate

[104] Robert Baldwin, Martin Cave, and Martin Lodge, *Understanding Regulation: Theory, Strategy, and Practice* (first published 1999, 2nd edn, OUP 2011) 343-44. See also section 1.2.

[105] Nathan Cortez, 'Regulating Disruptive Innovation' (2014) 29 Berkeley Technology Law Journal 175, 203-04.

[106] Mark R Patterson, *Antitrust Law in the New Economy* (Harvard UP 2017).

> interests exercise a kind of hydraulic pressure which makes what previously was clear seem doubtful, and before which even well settled principles of law will bend.[107]

Will AI exert 'hydraulic pressure' on settled norms? Again, courts have — for the most part — been able to adapt. In the absence of new forms of legal personality,[108] and presuming that conduct by AI systems can be attributed to traditional legal persons and that evidentiary burdens can be met, the problems of speed, autonomy, and opacity pose difficult but not insurmountable challenges.

For the most part. On the margins, as we have also seen, AI systems create risks or enable conduct that does not fall neatly into existing categories. Though enterprising judges will endeavour to apply laws sensibly, even as agencies and legislatures strive to ensure the relevance of those laws and their implementation, it would be prudent to add an entity tasked precisely with the function of identifying and addressing those gaps as they arise.

3.4 An AI Ombudsperson?

Though various jurisdictions have long had comparable officials, the term ombudsperson (or ombudsman) has Scandinavian roots. In general, it refers to an individual appointed by the state to represent the interests of the public. He or she typically enjoys some measure of independence and flexibility in his or her mandate, which is sometimes cast as upholding administrative justice, human rights, or the rule of law itself. In addition to responding to complaints, that mandate may extend to representing the public interest with respect to systemic issues.[109]

Powers of enforcement may be limited — classically, an ombudsperson was limited to 'soft' powers of investigation, recommendation, and reporting. Despite these limitations, ombudsperson institutions were embraced as a tool to address diverse accountability

[107] *Northern Securities Company v United States*, 193 US 197, 400-01 (1904).

[108] See Simon Chesterman, 'Artificial Intelligence and the Limits of Legal Personality' (2020) 69 International and Comparative Law Quarterly 819.

[109] See Varda Bondy and Margaret Doyle, 'What's in a Name? A Discussion Paper on Ombud Terminology' in Marc Hertogh and Richard Kirkham (eds), *Research Handbook on the Ombudsman* (Edward Elgar 2018) 485; Richard Kirkham and Chris Gill (eds), *A Manifesto for Ombudsman Reform* (Palgrave Macmillan 2020). Cf Lord Sales, 'Algorithms, Artificial Intelligence, and the Law' (2020) 25 Judicial Review 46, 54-57 (proposing an expert algorithm commission).

problems in the latter half of the twentieth century as 'ombudsmania' took hold. In the 1980s this overlapped with human rights discourses; from the mid-1990s it was linked to global governance. Today, the International Ombudsman Institute boasts member institutions in more than 120 countries.[110]

Though many such offices have mandates that cut across the public sector or beyond, dedicated ombudsperson institutions have proven useful in other areas where traditional regulation is inadequate. In relation to national security concerns, for example, the ability to address complaints with a degree of informality has on occasion been more effective than judicial processes.[111]

In some countries the term commissioner, inspector-general, or people's advocate may be preferred. The precise name is less important than the office's independence, mandate, powers, and resources. Independence from government and industry is essential if it is to be taken seriously. In addition to avoiding regulatory capture, this should assist in being able to cut across administrative siloes. The mandate should be framed broadly as identifying and addressing harms and injustice caused by AI that cannot be prevented or resolved through existing norms and institutions. This should include the ability to initiate inquiries as well as respond to complaints. (Limiting transparency to explainability, for example, puts an undue onus on individuals to *know* that they have been harmed and initiate an inquiry themselves.[112])

To be effective, the ombudsperson needs to be able to require cooperation and have access to relevant documents, including those that would otherwise be privileged. Though proceedings can be confidential, it is vital that there be an option to make the outcome public. Reports should not be limited to resolving disputes but should include the ability to make wider recommendations to change practices, policies, and legislation. Those

[110] Charles S Ascher, 'The Grievance Man or Ombudsmania' (1967) 27 Public Administration Review 174; Chris Gill, 'The Evolving Role of the Ombudsman: A Conceptual and Constitutional Analysis of the "Scottish Solution" to Administrative Justice' [2014] Public Law 662; Tero Erkkilä, *Ombudsman as a Global Institution: Transnational Governance and Accountability* (Palgrave Macmillan 2020).

[111] Simon Chesterman, *One Nation Under Surveillance: A New Social Contract to Defend Freedom Without Sacrificing Liberty* (OUP 2011) 218.

[112] Cf Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law & Technology Review 18, 83-84.

recommendations need not be binding, but best practice is for the legislature or other receiving body to be required to give reasons for not accepting them.[113]

Much of the work of an AI Ombudsperson might be redirecting cases to appropriate government agencies or the relevant part of the legal system. Yet the role should go beyond ensuring legality and compliance: the value of an ombudsperson is in promoting human rights and good administration.[114] In the European Union, Data Protection Authorities fulfil some of these functions.[115] They might also be taken on by existing ombudsperson institutions. Indeed, in March 2020 the International Ombudsman Institute organized a workshop with Catalan's Ombudsman on the role of ombudsperson institutions in protecting and upholding human rights in a world of AI.[116] Given the steep learning curve and the likely expansion of the impact of AI, however, a dedicated office — either standalone or as part of a larger entity — would give the issue the proper attention and prevent wheels being constantly reinvented.

4 Conclusion

One consequence of Eisenhower's 'Atoms for Peace' speech was the biggest scientific conference the world had ever seen. Proposed by the United States and convened by the General Assembly in 1955, the First International Conference on the Peaceful Uses of Atomic Energy, later known as the First Geneva Conference, brought together some 1,500 delegates from 38 countries, with over 1,000 papers presented. The Second Geneva Conference, held in 1958, was nearly twice as large. It was a period of euphoria and optimism, with many states establishing nuclear research and development programmes even as the IAEA Statute was being drafted and ratified.[117]

The limitations of an analogy between nuclear energy and AI are obvious. Nuclear energy refers to a well-defined set of processes related to specific materials that are unevenly

[113] Developing and Reforming Ombudsman Institutions (International Ombudsman Institute, June 2017).

[114] P Nikiforos Diamandouros, 'From Maladministration to Good Administration: Retrospective Reflections on a Ten-Year Journey' in Herwig CH Hofmann and Jacques Ziller (eds), *Accountability in the EU: The Role of the European Ombudsman* (Edward Elgar 2017) 217.

[115] General Data Protection Regulation 2016/679 (GDPR) 2016 (EU), art 57.

[116] Ombudsmen Alert About Artificial Intelligence and Human Rights (International Ombudsman Institute, 11 March 2020).

[117] Robert A Charpie, 'The Geneva Conference' (1955) 193(4) Scientific American 27; Fischer (n 76) 31.

distributed; AI is an amorphous term and its applications are extremely wide. The IAEA's grand bargain focused on weapons that are expensive to build and difficult to hide; weaponization of AI promises to be neither.

Nonetheless, some kind of mechanism at the global level is essential if regulation of AI is going to be effective. This article has argued that industry standards will be important for managing risk and states will be a vital part of enforcement, with gaps to be plugged by an AI Ombudsperson or equivalent institution at the national level. In an interconnected world, however, regulation premised on the sovereignty of territorially-bound states is not fit for purpose. The hypothetical IAIA offered here is one way of addressing that structural problem.

Yet the biggest difference between attempts to control nuclear power in the 1950s and AI today is that when Eisenhower addressed the United Nations, the effects of the nuclear blasts on Hiroshima and Nagasaki were still being felt.[118] The 'dread secret' of those weapons, he warned, was no longer confined to the United States. The Soviet Union had tested its own devices and the knowledge was likely to be shared by others — perhaps all others. Doing nothing was to accept the hopeless finality that 'two atomic colossi are doomed malevolently to eye each other indefinitely across a trembling world'.[119]

There is no such threat from AI at present and certainly no comparably visceral evidence of its destructive power. Absent that threat, getting agreement on meaningful regulation of AI at the global level will be difficult. One reason why the UN Security Council enjoys powers that its predecessor in the League of Nations lacked is that the member states negotiated the UN Charter while the bombs of the Second World War were still falling. The final document was crafted in aspirational but knowing language, promising to save succeeding generations from 'the scourge of war, which twice in our lifetime has brought untold sorrow to mankind'.[120]

It is conceivable that AI itself will help solve the problems raised here. If it does not, global institutions that might have prevented the first true AI emergency will need to be created in a hurry if they are to prevent the second.

[118] Lesley MM Blume, *Fallout: The Hiroshima Cover-Up and the Reporter Who Revealed It to the World* (Scribe 2020).

[119] Atoms for Peace (n 6).

[120] UN Charter, preamble.